

Secrecy damages the NIH

An extensive piece of investigative journalism has highlighted conflicts of interest that cast a pall over the National Institutes of Health. The agency will lose its well-earned public trust if it does not radically increase its transparency.

At first sight, it looked like the biggest misconduct scandal for years: scientists at the world's largest research campus riding the gravy train of lucrative private consulting deals that have distorted research and possibly even led to the deaths of patients in clinical trials. At least, that is how an extensive article in the *Los Angeles Times* on 7 December portrayed the situation at the US National Institutes of Health (NIH).

In response to the coverage, a congressional committee has demanded records of consulting arrangements at the NIH going back four years. Hearings seem likely. And NIH director Elias Zerhouni has promised a thorough review of consulting practices and ethics rules.

Although such responses are appropriate, the magnitude of the offences seems unlikely to be as worrying as the article implied — which is probably why other media have not pounced on the story. What appears at first to be damning evidence of industry distortion of public medical research is actually rather inconclusive. The article details the consulting deals of a number of high-ranking NIH officials and describes how each posed an alleged conflict of interest. But it is also possible that each official was behaving within ethical bounds.

In one example, Stephen Katz, director of the National Institute of Arthritis and Musculoskeletal and Skin Diseases, was reportedly reluctant to halt a trial of an anti-inflammatory drug in the wake of a patient's death, while at the same time serving as a paid consultant to the maker of the drug. But Katz says he did not know that the drug was made by his client, because it was provided by a differently named

US affiliate. Furthermore, as with all the other consulting deals mentioned in the article, Katz's was subject to ethical review and approval.

Permitting employees to engage in private consulting is essential to the vitality of the NIH. By the mid-1990s, the agency was losing talent to industry and academia, where higher pay and private consulting are common. When Harold Varmus took the helm in 1993, he was charged with revitalizing research. By allowing researchers to increase their consulting income, he helped to turn the tide.

The NIH's policies on such matters are open to public scrutiny (see <http://ethics.od.nih.gov>), but the practice is not. Private consulting, which in the NIH must be done in "personal time", can open the way for conflicts of interest. Even assuming that proper safeguards are in place, one disturbing fact remains. Of the 2,259 current consulting deals made by NIH employees in the upper salary brackets, only 127 are subject to full public disclosure — a statistic that even Varmus says he finds disconcerting. The rest fall through an administrative loop-hole that permits many top-level researchers to file minimal details of their outside agreements, and then for internal use only.

As the pending congressional review demonstrates, the appearance of impropriety can be as damaging as impropriety itself. The only way forward is transparency. Public officials throughout government engage in private consulting, but that consulting must be open to public scrutiny. The public should not be asked to take the NIH's word that its staff are behaving well in such sensitive areas of responsibility. They should be able to see for themselves. ■

Variation for all

The consortium that is mapping human haplotypes establishes some important principles of access and credit in this issue.

What is the logical next step after completing the sequence of one human's genome? The international human genomics community thinks it is to catalogue the differences between humans, and to decode their importance to health and diseases. So researchers are building a map of common variation in the human genome, using combinations of markers that yield mosaic-like patterns called haplotypes (see *Nature* **421**, 409–412; 2003).

The International HapMap Project (www.hapmap.org; short for haplotype mapping) is a large-scale public and private research collaboration that will use 270 human genomes from four populations to generate a publicly available tool for use by biomedical researchers anywhere (see *Nature* **425**, 758–759; 2003). Given the project's pledge to ensure rapid and complete data release, a question arises about how those doing the work receive the credit for their contributions.

In January 2003, representatives of the publicly funded genomics community met in Fort Lauderdale, Florida, to address the tension between public access to large-scale data sets and proper credit for the researchers (see *Nature* **421**, 875; 2003 and www.wellcome.ac.uk/en/1/awtpubrepdat.html). A proposed solution was to define certain projects as 'community resource projects', where different rules apply for the distribution of data pre-publication and for publication. Examples include the recent publicly funded human and mouse

genome-sequencing projects, and the International HapMap Project. For community-resource projects, the group suggested a new type of publication called a 'project description', the purpose of which is "to inform the scientific community about the resource project and to provide a citation to reference the source of the data".

We are pleased to publish, on page 789 of this issue, the project description for the International HapMap Project. Anyone who uses the HapMap data in their own publication — even before those who generated the primary data have published their analysis — can now properly cite the plans laid out by the originators. Another benefit is that the many scientists at all levels who are already contributing to an important project have something to show for it in their CVs.

In the paper, the International HapMap Consortium describes its strategy for dealing with patient consent and ascertaining samples, and the inherent conflict in making data available rapidly and freely while heading off possible patenting issues by imposing a 'click-wrap' agreement on individual genotypes. As with many large-scale projects, the process leading to the results is complex, and issues such as the best way to analyse the mountains of data remain to be resolved. *Nature* supports this publication of the planned way forward to generate discussion, to give credit where it's due, and to promote the idea that such openness is in the best interests of science. ■





A shot in the arm

Experts seek fresh approach to flu vaccines
p742



Sanity clause

Cloned cattle show resistance to mad cow disease
p743



Meet the family

Coral study hints at lengthy genetic lineage
p744



Ape measure

Chimp genome offers clues to what makes us human
p745

Accusations of bias prompt NIH review of ethical guidelines

Jonathan Knight, San Francisco

The US National Institutes of Health (NIH) is to reassess its ethics rules following allegations that consulting fees paid to some of its top scientists have biased decisions in clinical research.

The charges, levelled by the *Los Angeles Times* on 7 December, centre on consulting fees paid to a number of the agency's top scientists by private drug firms. The paper alleged that such deals had influenced research decisions involving the companies' products.

It also suggested that the problem runs deep and is hidden from public view by rules that allow most NIH scientists to keep their consulting deals confidential.

The office of NIH director Elias Zerhouni issued a statement on 10 December to say that the allegations are being investigated but that all employees are believed to have followed government ethics rules. Nevertheless, it added: "We will need to consider changes after a thoughtful analysis of the issue."

The House Committee on Energy and



Are consultancy fees exerting undue influence on clinical trials?

Commerce, which oversees the NIH, has demanded complete records of all consulting deals made by in-house NIH scientists since 1999. The agency has offered its full cooperation with the investigation.

The NIH campus in Bethesda, Maryland, is the world's largest biomedical research centre. But until 1995, scientists there faced strict limits on their consulting activities. These rules were revoked by then director Harold Varmus

in an attempt to make the NIH more attractive to top researchers. Under the new rules, scientists continued to be barred from consulting if they were also directly involved in a company's dealings with the NIH, and all deals were subject to approval by ethics staff.

But according to the *Los Angeles Times*, these guidelines have not prevented conflicts of interest — and in some instances may have biased decisions on clinical research. Cecil Fox, a private research consultant in Little Rock, Arkansas, says that the paper's article is right on target. Fox was a scientist at the National Institute for Allergy and Infectious Diseases from 1973 to 1991, and consulted for many labs at the Bethesda campus until last year. He says that private deals with industry have seriously compromised the independence of many NIH researchers.

Colleagues of the researchers named in the *Los Angeles Times* article have been quick to spring to their defence. One, who spoke on the condition of anonymity, said that the ethics charges were completely overblown. "I know the people and they are people of the highest integrity," he said.

But the high degree of confidentiality surrounding consulting deals at the NIH is disturbing even some of the agency's defenders. The 1989 Ethics Reform Act requires prominent or highly paid government employees to file public disclosures of outside income. Other employees, meanwhile, file confidential reports that can be reviewed by Congress, but that are not available to the public.

With more than 94% of its high-earning employees falling into the bracket, the NIH has one of the highest percentages of confidential filers in the government. In part, this is because the agency has widened its pay bands in recent years, assigning highly paid staff to grades that allow confidential filing.

These wider pay brackets were the only way the NIH could increase the salaries of high-level scientists that it didn't want to lose, says Varmus. But he adds there should be greater public transparency from the NIH: "I think that should be reviewed at this point."

Scientists attack industrial influence

Jonathan Knight, San Francisco

Four scientists who feel that their work has been suppressed by the biotechnology industry issued a call for greater scientific freedom at a rally on 10 December at the University of California, Berkeley.

Nearly 500 people turned up to hear the researchers describe the obstacles they claim to have faced in disseminating research findings that questioned the safety of pesticides and genetically engineered crops.

"I want to make this the beginning of something," said Ignacio Chapela, the Berkeley ecologist who organized the event. Chapela claims that he was denied tenure at Berkeley last month because of his outspoken stance against corporate financing of public research (see *Nature* 426, 591; 2003).

At the rally, Berkeley amphibian

biologist Tyrone Hayes gave a detailed account of what he called industry attempts to suppress and discredit his work on atrazine, the most widely applied herbicide in the United States.

Joining the rally on an Internet video link was John Losey, an entomologist at Cornell University in Ithaca, New York, who in 1999 found that pollen from insecticide-producing maize is toxic to monarch butterfly larvae. Losey said he had been asked by industry representatives not to publish his findings.

One sceptic asked if the panellists were really victims of suppression, when their findings had attracted so much publicity. Chapela responded that for each panel member there were many others whose careers were ended by industry pressure. "We are the survivors," he said. ■

Disillusionment and doubt undermine Kyoto's birthday bash

Quirin Schiermeier, Munich

It was a rather forlorn party that took place last Thursday night at the Milan Conference Centre. On the penultimate day of the United Nations' climate talks, hundreds of delegates commemorated the sixth birthday of the Kyoto Protocol on Climate Change — but no one was quite sure if it was a jubilee or a funeral.

The subdued atmosphere at the ninth conference of parties to the Kyoto Protocol could be attributed in large part to rampant speculation that Russia may not ratify the treaty, meaning that it will not come into force (see page 756).

Partygoers no doubt also had their minds on the many alarming reports released at the meeting. The World Health Organization, for example, reported that climate change seems to be responsible for some 150,000 deaths each year — a figure that is expected to double by 2030. Another group asserted that a mere 2 °C of warming over pre-industrial temperatures would be "intolerable" for the planet (see *Nature* 426, 486; 2003).

But there was some Christmas cheer for those in Milan. Some technical details of the protocol were finalized, such as a scheme in which industrial countries can acquire emissions credits for financing land-use projects that absorb carbon.

The European Union also pledged €400 million (US\$490 million) per year from 2005 to help poorer countries adapt to climate change by improving flood protection or irrigation, for example.

Non-governmental organizations applauded Europe's efforts but criticized the US delegation, which argued in favour of carbon sequestration over emissions cuts. "There is no choice between future technology and taking action now. We need both," says Alden Meyer, director of policy at the Union of Concerned Scientists in Washington DC. ■



'Reverse genetics' could offer forward-thinking flu vaccine

Helen Pearson, New York

Health authorities are under renewed pressure to modernize influenza-vaccine production after an early surge in flu has highlighted shortfalls in the current jab.

Concerns about the vaccine were raised after an early and sudden start to the influenza season in Europe and the United States in late October. Public-health experts are still not sure whether the number of cases will go down — or rise to epidemic levels.

The spike in cases has highlighted two problems with the widely used inactivated influenza vaccine. First, the three virus strains in the vaccine exclude this season's predominant strain, a version of the A(H3N2) influenza subtype called Fujian. Public-health officials say that the closely related Panama strain in the vaccine does confer some protection against the Fujian strain.

Second, parts of the United States have run short of the inactivated vaccine, for which the virus is mass-produced in hens' eggs. This has prompted the Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia, to look for remaining stocks from US and foreign vaccine manufacturers.

Influenza experts at the World Health Organization (WHO) say that they knew the Fujian strain was circulating when they met in February 2003 to decide which strains to include in this season's Northern Hemisphere shot. The vaccine is regularly updated because the flu virus rapidly mutates to dodge our immune response. It shuffles two genes that encode proteins on the virus's coat, called haemagglutinin and neuraminidase.

But laboratories in the WHO's network failed to find a Fujian strain that would grow in fertilized hens' eggs in time to include in this season's vaccine. To make a suitable vaccine strain, researchers inject the circulating virus, such as Fujian, and another, fast-growing flu strain into eggs, where the two mix and match their genes.

From the eggs, scientists aim to pull out a new fast-growing reassorted virus — called a seed strain — which carries the haemagglutinin and neuraminidase genes from Fujian. Vaccine manufacturers need to receive the seed strain in time to grow it up in tens of millions of eggs, a process that is proven, efficient and reliable, but takes up to six months. "Using eggs does seem a bit like ancient technology," admits Alan Hay, who heads the WHO influenza collaborating centre in London.

Experts say that problems growing the Fujian strain could be bypassed using a technique called reverse genetics. Scientists could



Feel the strain: publicly available flu jabs may not protect against this winter's predominant virus.

rapidly engineer the seed strain in the laboratory by stitching together the viral genes they want, and then use it to mass-produce the vaccine in hens' eggs.

Influenza vaccines made using reverse genetics have not yet progressed through clinical trials, however. And according to WHO officials and other experts, vaccine manufacturers might be dissuaded from using reverse genetics because they would owe licensing fees to MedImmune, a biotechnology firm based in Gaithersburg, Maryland that holds patents on the technique.

The WHO will organize a series of meetings between vaccine manufacturers, regulatory authorities and other parties in Geneva, Switzerland, in February 2004 to explore these issues, says Klaus Stohr, head of its influenza programme. The WHO is also working with the US National Institutes of Health to push ahead a clinical trial of a vaccine made using reverse genetics, which is expected to start in late 2004.

Experts say that flu vaccine could also be mass-produced in mammalian cell cultures, which could be set up for production in two weeks, whereas eggs have to be ordered at least six months in advance. "It can buy you some time," says Bram Palache at Solvay Pharmaceuticals, which is developing the technology at Weesp in the Netherlands.

Manufacturers have yet to prove that vaccines made in cell culture are as cheap as egg-made vaccines, but Palache predicts that they will be licensed and available in Europe by 2005. And while updated techniques cannot make up for the vaccine shortfall this season, experts are keen to get them in place in time to combat future, more aggressive flu strains. "We are concerned that it be discussed before the pandemic," says Stohr. ■

Koreans rustle up madness-resistant cows

David Cyranoski, Tokyo

Cloned cattle that are said to be resistant to mad cow disease have been born in Korea. Whether milk or meat from such cows will ever reach supermarkets is unclear. But some experts are enthusiastic about the prospects of using such animals as 'bioreactors' to produce drugs for human use.

Mad cow disease, or bovine spongiform encephalopathy (BSE), is thought to be caused by prions, aberrant forms of a normal protein called PrP. Prions convert other PrP to their own misshapen form, causing plaques of protein to accumulate, killing nearby nerve cells. More than 130 people, most of them in Britain, have died from BSE since the disease was found to have crossed over to humans in 1996.

Today, officials in many countries keep BSE out of the human food chain by monitoring slaughterhouses and cattle for signs of infection. But Korean researchers led by Woo-Suk Hwang, a cloning specialist at Seoul National University's College of Veterinary Medicine, claim that their calves could make BSE a thing of the past. On 10 December, they announced the birth in November of four 'BSE-proof' cloned calves.

The clones were made from a cell carrying additional genes for PrP that had been mutated so that although the protein still binds to prions, it resists being converted to the aberrant form. Hwang's team also introduced promoter DNA that boosts the production of this altered PrP well above normal levels, minimizing the chance that prions can find and convert normal PrP. A similar method has been used previously in mice (V. Perrier *et al. Proc. Natl Acad. Sci. USA* **99**, 13079–13084; 2002), where it does seem to protect against prion infection.



Beefed up: Korean researchers believe tests will show that their cloned cows are resistant to BSE.

Hwang says that more testing is needed. From February, the calves will be sent to a new BSE facility in Tsukuba, Japan, where they will be given feed containing BSE-inducing prions to see if they are truly immune. The group will also test another group of cloned cattle, due to be born in January, in which the gene for PrP has been knocked out altogether, giving prions nowhere to get a foothold.

"If we have positive results after three to five years we will try to make these cattle commercially available," says Hwang. Last week, the group applied for a global patent covering both methods of generating the clones.

The tests must look at whether the treatments affect the animals' health. The function

of PrP is unknown, and knockout mice seem to be healthy. "But the protein's function might be different in mice and cattle," warns Eckhard Wolf, who studies transgenic animals at the University of Munich, Germany.

Wolf doubts that transgenic BSE-proof cows will be important for agriculture, given the cost of producing transgenic animals and the huge investment in conventionally bred cattle. But he does see an opportunity to create cows engineered to produce therapeutic proteins in their milk. Fears about BSE could deter regulators from approving such drugs for use, Wolf suggests. "Transgenic cows that cannot be infected with BSE should be clinically more acceptable," he says. ■

EMBO chief threatens to quit over funding crisis

Quirin Schiermeier, Munich

A freeze on funding for the European Molecular Biology Organization (EMBO) has prompted its executive director, Frank Gannon, to threaten resignation unless prospects improve.

The organization promotes molecular biology through postdoctoral fellowships, conferences and workshops, and is funded by the 24 member countries of the European Molecular Biology Conference.

This summer, the conference unanimously accepted a five-year plan outlining EMBO's funding priorities for 2004–08. The plan includes an extension of EMBO's two-year fellowship programme to three years. This programme supports the training of young postdoctoral biologists

and allows them to travel between labs.

But it now seems that there will not be enough money to pull such plans off. Last month, the conference's four largest members — Germany, Britain, France and Italy — unexpectedly rejected any increase in EMBO's budget for 2004–06.

The situation is all the more worrying, says Gannon, because the number of applicants for the fellowship programme has leapt in the past two years, probably because of the poor job market for biologists in countries such as Germany, France and Spain. But as the number of applicants has risen, the money to support them remains the same. The success rate for applicants has fallen from 22% before 2001 to below 15% now, and may continue to fall.

"I would not be able to stand over an inadequately funded programme, as it would damage EMBO's reputation," says Gannon.

EMBO asked conference members for a budget increase from €10.8 million (US\$13.2 million) to €15.5 million in 2006. The contribution of different countries is linked to economic performance, so Britain would shoulder the biggest rise, of some €1 million by 2006 — a 44% increase.

"We didn't say that we couldn't agree to any increase, but we felt that 44% is too much," says David Smith, an official at the UK Medical Research Council, which funds Britain's membership of the conference.

Conference members will discuss the possibility of obtaining further funding at a special meeting in January. ■

Coral reveals ancient origins of human genes

Carina Dennis, Sydney

A study of coral suggests that ancient members of the animal kingdom slithered through the Precambrian mud with a hefty cache of genes in common with humans.

Surprisingly, many of these genes are not shared with creatures such as flies and worms, even though these animals evolved millions of years after coral. This calls into question some studies that use these model organisms to unravel the evolution of the human genome.

The investigation, published in this week's *Current Biology* (D. Kortschak *et al.* 13, 2190–2195; 2003), looked at some 1,300 gene sequences expressed in the coral *Acropora millepora*, and found that about 500

sequences had matches in gene databases. These sequences, called expressed sequence tags, represent either single genes, different pieces of the same gene, or expressed portions of DNA that do not contribute to a coding gene.

Of these, 90% were present in humans, and about 10% were found in humans but not in the fruitfly *Drosophila melanogaster* or the nematode worm *Caenorhabditis elegans*. This finding suggests that many genes thought to be vertebrate-specific may in fact have much older origins, and have been lost during the evolution of the fly and worm.

"The assumption was that coral would lack many of the genes found in higher animals," says Robert Saint of the Australian

National University in Canberra, one of the study's authors. Instead, they were surprised to find genes similar to those that contribute to the specialized tissues of vertebrate nervous systems, even though coral has only a simple nerve net.

"There are important basic scientific questions that we need to ask using the coral system that should tell us about the evolution of developmental mechanisms in the animal kingdom," says David Miller, a molecular biologist at James Cook University in Townsville, Australia, and co-author of the study.

The idea that genes previously regarded as 'vertebrate innovations' may have evolved before vertebrates did isn't new. Alejandro Sánchez Alvarado, a developmental biologist at the University of Utah, previously found genes in the flatworm *Schmidtea mediterranea* that were thought to have evolved in vertebrates (A. S. Alvarado *et al. Development* 129, 5659–5665; 2002).

But the idea that some animals may discard genes as they become more sophisticated is still controversial. "We won't really know until we have more worm and insect genomes to compare," says developmental biologist Eric Davidson of the California Institute of Technology in Pasadena.

The finding means that although fly and worm models are useful for studying gene function in development and cellular processes, they may be of limited value in studies of the evolution of human genes. "We need to look at many other animal genomes that haven't undergone the same degree of gene loss to understand the evolution and function of human genes, and how they generate complexity," says Sánchez Alvarado. ■



The coral *Acropora millepora* shares a surprisingly large number of genes with vertebrates.

Elsevier waves goodbye to BioMedNet web portal

Jim Giles, London

The popular life-sciences website BioMedNet is to close down for good. Observers say that the decision by scientific publisher Elsevier, which owns it, heralds a move away from general scientific websites, towards more specialized services.

Elsevier bought BioMedNet in 1997 from Vitek Tracz, an online-publishing pioneer who developed it for the London-based Current Science Group, which he chairs. BioMedNet evolved to bring together several Elsevier publications, such as the Trends series of journals, with conference reports, information on lab equipment, and access to the Medline database of life-sciences papers.

The website's 12 staff in London were told on 1 December that Elsevier was pulling the plug on BioMedNet, together with other

general-purpose websites, including Chemweb and ElsevierEngineering.

"Having carefully reviewed the options available to us, we have decided that investments in the science and technology portals BioMedNet, Chemweb and ElsevierEngineering.com will be withdrawn," says a spokesperson for Elsevier.

When Elsevier bought BioMedNet, many experts anticipated that websites would soon generate substantial advertising revenue. "The web was hot," recalls Tracz. "People were full of expectation."

But David Worlock, chairman of London-based publishing consultancy EPS, says that Elsevier tested several different Internet business models, such as ScienceDirect, a website that provides access and search functions for Elsevier journals and reference

works, but has no other content of its own.

Worlock says that most scientists now go straight to journal articles using sites such as ScienceDirect, rather than via portal services such as BioMedNet. Elsevier is also locked in tough subscription negotiations with university libraries (see *Nature* 426, 217; 2003), sparking suspicions that revenue fears may have influenced their decision, he says.

Elsevier is expected to focus its Internet efforts on sophisticated search tools to complement ScienceDirect. The company is said to be developing an application known as Scopus, which will allow users to probe the network of citations that link related papers. Such a service would be a rival for ISI Web of Science, a tool marketed by Philadelphia-based Thomson ISI. ■

▶ www.bmn.com



Close cousins: chimps share much of our DNA.

Chimp genome hints at little differences that make us human

Washington A team of scientists has put together a draft of the genetic sequence of our closest relative, the chimpanzee *Pan troglodytes*. The sequence, which was stitched together in less than a year by a group led by Eric Lander at the Broad Institute in Cambridge, Massachusetts, and Richard Wilson at Washington University School of Medicine in St Louis, Missouri, is around 90% complete.

Researchers keen to compare the genomes of different species (see page 750)

have already pounced on the draft. Michele Cargill of biotech company Celera Diagnostics in Alameda, California, and her colleagues have compared more than 7,500 chimpanzee, human and mouse genes (A. G. Clark *et al. Science* **302**, 1960–1963; 2003). Their results indicate that about 1,500 genes have been affected by evolution since chimps and mankind went their separate ways around five million years ago. Of those human genes with a known function — about half of the total — nearly 50 are linked to smell, 21 to hearing, and 80 to digestion of proteins, perhaps highlighting some of the main ways in which the species differ.

♦ www.ncbi.nih.gov/Genbank

Nobel winner helps particle project get back on beam

Tokyo A Japanese neutrino project is back on track this month, after a Nobel laureate said he failed to understand why it derailed.

The Next Generation Long-Baseline Neutrino Oscillation Experiment is set to be built in Tokai. But the Japanese Council for Science and Technology Policy gave the project the lowest possible rank in its priority listings this October, saying that it could not justify the price tag of ¥600 billion (US\$5.6 billion) for construction and

operation over 20 years. The decision looked set to delay or even cancel the project.

Last month, Masatoshi Koshiha, who won the 2002 Nobel Prize in Physics for his work on neutrinos, told the council's members that such a low status for the project was "inconceivable". Shortly afterwards, the council notified the finance ministry that the low priority ranking was no longer valid, saying that a follow-up report from the education ministry had clarified the importance of the project.

The neutrino beam is now likely to receive funding in December, and construction should begin next April.

♦ www-nu.kek.jp/jhfnu/index_e.html

Italy takes hard line on embryo research

Munich A law approved by the Italian Senate last week, which makes Italy one of the most restrictive countries in Europe for assisted reproduction, has prompted an uproar from researchers and clinicians.

The proposed law, which awaits the formality of final approval by the Italian Chamber of Deputies in January, bans research on human embryos for any purpose, including extracting stem cells for research. It also means that doctors will not be able to freeze or store fertilized eggs for

use in *in vitro* fertilization, nor will they be able to test embryos for genetic diseases before implanting them. Moreover, single women will not be eligible for fertility treatment, and the use of anonymous sperm donations to sperm banks will be forbidden. Those who fall foul of the law face prison terms and fines of up to $\approx$ 1 million (US\$1.2 million).

There is a loop-hole for researchers, however — the law doesn't explicitly prevent Italian scientists from importing and using stem-cell lines created at any time from spare embryos in other countries.

Japan sees red-planet mission end in failure

Tokyo Japan's space agency has given up on its mission to Mars. The Nozomi probe was originally due to arrive at the red planet 15 months after its launch in July 1998. But in December that year it suffered engine trouble, forcing the agency to reschedule its arrival for 2003. Last week, Japan's space agency announced that it has abandoned last-ditch efforts to direct the craft into a martian orbit.

Instead, the agency has diverted Nozomi into a solar orbit to prevent it crashing into Mars — a collision that some fear would have contaminated the planet with microbes from Earth. Nozomi will not be able to send

back data on solar activity from its vantage point, however, as some of its equipment was damaged by solar flares in the spring.

Volcano breathed life into Munch's *Scream*

Washington A volcanic eruption in Indonesia that turned sunsets blood red around the world was the likely inspiration for Edvard Munch's *The Scream*, says a report in the February 2004 issue of *Sky and Telescope*.

Donald Olson, an astronomer at Texas State University in San Marcos, and his team tracked down the precise location of the



The Scream: influenced by Krakatoa's eruptions.

painting. They determined that Munch was facing southwest when composing the piece — the direction of the 'Krakatoa twilights' that appeared in European winter skies in 1883–84 (see *Nature* 29, 130–133; 1883).

Britain launches inquiry into cost of science journals

London Britain's parliamentary Science and Technology Committee has launched an inquiry into scientific publishing.

The committee, chaired by Ian Gibson, Labour member for Norwich North, announced the inquiry on 10 December in response to concerns that libraries are unable to afford expensive journals (see *Nature* 426, 217; 2003). The inquiry will look at what the government can do to make the market more competitive. It will also examine the effects of open-access publishing (see *Nature* 425, 554–555; 2003), and investigate whether the government should support it.

"Researchers, teachers and students must have easy access to scientific publications at a fair price," says Gibson. "The committee will have some very tough questions for publishers, libraries and government on these issues." The committee is now accepting written evidence and will hold discussion groups in March. Its final report is scheduled for next summer.

2003

IN CONTEXT

It was a year with more than its fair share of unsettling events. Impressive advances, such as a high-resolution microwave map of the Universe, were tempered by war, SARS and the loss of the space shuttle Columbia. *Nature's* reporters explore the highs and lows of 2003.

Climate of conflict

In the shadow of war

No one knows exactly what was going through the mind of David Kelly, an expert on bioweapons employed by the UK Ministry of Defence, on the afternoon of 17 July. All we know is that this softly spoken microbiologist had been driven over the edge, following his identification as the source of a now infamous BBC radio story. This item alleged that the British government “sexed up” a dossier detailing Saddam Hussein’s weapons of mass destruction to justify the attack on Iraq. Police found Kelly’s body at 9.20am on 18 July, in woods near his Oxfordshire home; his left wrist had been slashed.

The events that led to Kelly’s presumed suicide have since been subjected to a full public inquiry. Its chair, the judge Lord Hutton, is now preparing his report, which will be published in January. The BBC, government officials and ministers are all expected to come in for heavy criticism. But for Kelly’s family, and his scientific colleagues, the report will do little to lessen the sense of loss.

Unfortunately, Kelly wasn’t the only scientist whose life was this year torn apart by the ongoing ‘war on terror’. On 1 December, Thomas Butler, a microbiologist at Texas Tech University in Lubbock, shook his head and fought back tears as he was convicted of defrauding his employer, and of illegally

sending samples of the plague bacterium to Tanzania. Butler was found innocent of most of the charges that first landed him in trouble — that he illegally imported plague samples from the same country, and that he lied to federal investigators in claiming that they had gone missing. But many scientists have been chilled to the bone by his prosecution. Butler’s case is widely perceived as simply a warning to scare biologists who might be tempted to pay mere lip service to regulations on the handling of potential bioweapons.

Clamp-down

The circumstances that engulfed Kelly and Butler were admittedly extreme. But this year’s climate of conflict is affecting science, and scientists, in myriad ways — especially in the United States. Thousands of researchers working on human, animal and plant diseases are now struggling to get to grips with red tape designed to prevent pathogens getting into the wrong hands. The funding landscape in the United States has been transformed, with hundreds of millions of dollars being poured into projects to counter the terrorist threat. And for scientists trying to enter the United States from countries other than its close allies, immigration requirements have become stifling.

If you look at the US federal science

budget, there is little doubt that this is a country on a war footing. Since the mailed anthrax attacks of October 2001, the National Institute of Allergy and Infectious Diseases (NIAID) has distributed some \$1.8 billion for projects in biodefence. And in January, the new Department of Homeland Security opened its doors. With an annual research budget of almost \$900 million, the department aims to develop countermeasures, such as sensors, to thwart terrorist attacks. No other country has responded to the threat posed by groups such as al-Qaeda by redirecting its research effort in quite such a fundamental way. Indeed, speak privately with senior European science-policy officials, and they’ll tell you that they regard the US reaction as over the top.

As you might expect, trying to ramp up a huge counter-terrorist research programme has entailed a few teething troubles. Funding from the Department of Homeland Security has suffered delays. The NIAID, meanwhile, came under criticism for diverting cash from projects on AIDS and other diseases to fund the development of an anthrax vaccine.

The US government’s research priorities are also being influenced in more subtle ways by the ongoing conflict. For President George Bush’s administration, becoming independent of the oil imported from Middle East



countries where support for al-Qaeda is strongest has become a long-term priority. In his State of the Union address in January, Bush announced a \$1.2-billion initiative to develop hydrogen-powered cars. The United States also this year re-entered the \$5-billion international ITER project to develop a prototype nuclear-fusion power plant.

But for US microbiologists, the main issue is not so much the new funding landscape, but rather how to cope with new regulations designed to prevent bioweapons proliferation. As Butler claimed in his defence, many scientists find these regulations bureaucratic and confusing. The main new provisions cover pathogens from a list of about 80 judged to be of interest to terrorists wanting to attack people, crops or livestock. Labs working on these agents must register with the Centers for Disease Control and Prevention or the US Department of Agriculture (USDA), submit themselves to background checks by the FBI and complete a mountain of paperwork.

Plant pathologists, who were used to working with minimal government scrutiny, got a particularly rude awakening. Many plant scientists were puzzled by the inclusion of certain organisms on the list, and by the omission of others that seemed at least as dangerous. "The list needs some adjustment,"

Nature drew up a set of guidelines designed to prevent the publication of research that would provide a 'cookbook' for would-be bioterrorists (see *Nature* **421**, 771; 2003). But the US government has yet to introduce policies on the issue. "My own guess is that it will require some minor crisis to clarify the situation, such as a dispute between the US government and some scientific journal about whether an article should be published," says Will Happer, a physicist at Princeton University in New Jersey, who has advised the US government on numerous issues relating to science and security.

Shut out

Heightened concerns about security have also caused problems for foreign scientists trying to visit the United States. A new rule enacted in July requires virtually every visitor coming to the country for work or study to undergo an interview with an official at a US embassy. Foreign scientists are also subjected to security reviews involving several government departments. For hundreds of researchers, especially those from the Middle East, the countries of the former Soviet Union and China, this has meant delays that can last for months. "The most frustrating part is the lack of any information,"

says Olexei Motrunich, a physicist who has worked in the United States for the past eight years, and who in September was supposed to take up a postdoctoral position at the University of California, Santa Barbara. But he has been shut out of the country since July, when he went home to his native Ukraine for what should have been a brief visit.

As well as hampering scientists seeking employment in the United States, immigration delays have disrupted scientific meetings. "We had a number of difficulties getting visas for our overseas participants," says Keith Ellis, a theorist at Fermilab in Batavia, Illinois, who

helped to organize the XXI International Symposium on Lepton and Photon Interactions at High Energies, held in August. All but one of the Chinese delegation were forced to cancel, and many Russian researchers were similarly unable to attend.

Looking to the future, US science advocates are worried about the knock-on effects of the money being spent on the occupation and rebuilding of Iraq. "Spending on the war has to have some impact," says Sam Rankin, who heads the Coalition for National Science Funding, a lobby group in Washington DC. "We are going to have to work pretty hard." ■

Geoff Brumfiel and Jonathan Knight

HIGHLIGHTS

Bon voyage

NASA's Voyager 1 spacecraft has left the Solar System — probably. Space scientists are still quibbling about whether the probe has or hasn't actually crossed the critical boundary, called the terminal shock layer, where the solar wind hits the interstellar medium. Either way, Voyager is now 13.5 billion kilometres away from us — about 90 times the distance between Earth and the Sun. It should reach the next nearest star in 40,000 years.

Meet the ancestors

Fossils found near the village of Herto in Ethiopia were revealed to be the oldest yet of modern *Homo sapiens*. The skulls — of two adults and a child who lived 160,000 years ago — support the idea that modern humans originated in Africa, and that Neanderthals were a branch of the human evolutionary tree that later went extinct. Cut marks on the skulls suggest that the first modern humans had some mortuary practices. And marks on nearby animal bones suggest that these people may have dined on hippopotamuses.

Condensed matter

Two teams of physicists cracked the problem of making Bose-Einstein condensates out of molecules rather than atoms. Such condensates are a strange state of matter in which the constituent particles occupy the same quantum state and so behave as a single particle. Atomic condensates were achieved in 1995, but their molecular equivalents proved a harder nut to crack. By cooling atoms with a laser and squeezing them together with magnetic fields, researchers managed to create loose molecular bonds — the molecules then collapsed into a condensate.

Look, no gonads!

Both eggs and sperm were produced in the lab from mouse embryonic stem cells. Sperm cells, which were grown in culture from male embryonic cells, were even used successfully to fertilize natural mouse eggs. The lab-made eggs were, intriguingly, grown from both female and male cells. The ability to grow eggs in a dish should provide a boost for studies of fertility and cloning — if they turn out to be normal.

Blow up

The origins of long γ -ray bursts — which last longer than a few seconds — were pinned down at last. These bursts are some of the most energetic events in the Universe. Astronomers witnessed the death of a massive star and recorded a γ -ray burst from the same event. The spectacular stellar show proved that some, if not all, long γ -ray bursts are associated with supernovae, the catastrophic explosions that end the lives of the largest stars. The origin of shorter γ -ray bursts remains a mystery.



On guard: the conflict in Iraq has dominated the year and brought unexpected pressures to bear on the scientific community.

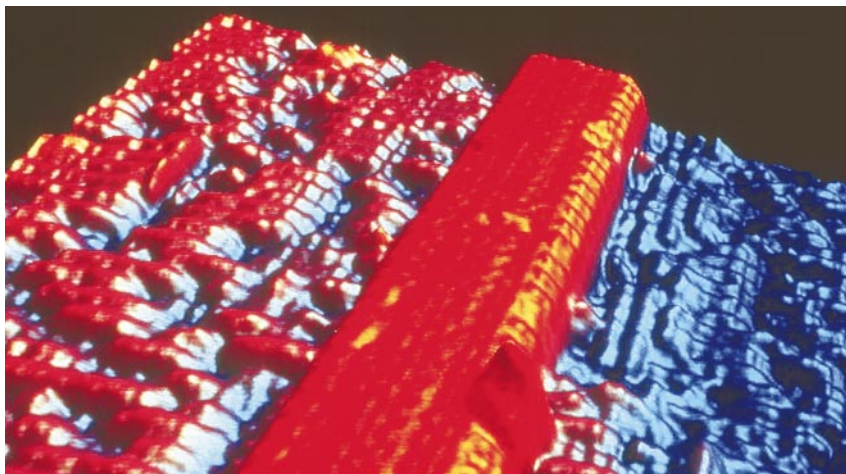
says Anne Vidaver of the University of Nebraska at Lincoln, who is serving on a committee of the American Phytopathological Society that is trying to develop objective criteria for placing agents on the list. Confusion was heightened in July by the sudden cancellation of a USDA meeting to discuss the list of restricted agents. No explanation was forthcoming, but some scientists believe that officials were worried about discussing information that might be useful to terrorists.

Similar confusion surrounds the question of whether results from certain projects should be published in the open literature. In February, a group of journals including

Nanotechnology

What is there to fear from something so small?

HEWLETT-PACKARD LABS/SPL



In miniature: nanowires such this one, just 10 atoms wide, are seen by many as the future of electronics.

Next March, Mark Welland's laboratory at the University of Cambridge, UK, will gain an unusual member of staff. Welland's team works on nanometre-dimension wires and tubes that could form the future of electronics, but the new recruit won't be an engineer or a physicist — he or she will be a social scientist.

The appointment — a two-year position that will include running regular meetings with everyone from industry representatives to green activists — acknowledges public fears about the possible effects of nanotechnology on human health and the environment. Although Welland may not subscribe to long-standing scare stories about a 'grey goo' of nanometre-sized robots taking over the planet, he realizes that scientists need to address this and other concerns head on.

Welland's proactive attitude stems in large part from a desire to avoid a rerun of Europe's experience with genetically modified (GM) crops. Over the past decade, vociferous campaigning by environmental groups, combined with an inability of both biotech firms and scientists to understand and respond to public concerns, has created a hostile climate of public opinion. Although the discussion has — in Britain, at least — matured this year into a more detailed scientific consideration of the environmental risks and benefits of individual GM crops, these negative perceptions are now deeply ingrained.

Given this background, Welland and other nanotechnologists are alarmed by signs that their work will be the next target for environmental campaigners. Media attention began to build after the publication in January of *The Big Down*, a report by the ETC group, a small environmental organization based in Winnipeg, Canada. In its previous guise as the Rural Advancement Foundation

International, ETC had a history of influential campaigning against GM agriculture.

The ETC report argued that synthetic nanoparticles could be toxic. *The Big Down* and other ETC publications went on to envisage a technology out of control, potentially able to create new weapons of mass destruction or novel life forms.

ETC doesn't subscribe to the old 'grey goo' scenario. But it has cleverly married fears about GM with its campaign against nanotech by arguing that the public should be especially concerned about 'nanobiotechnology'. This future industry, ETC claims, could create a 'green goo' of novel organisms and products that will behave in unpredictable and uncontrollable ways.

Ask ETC officials to define what, exactly, they mean by nanobiotechnology and they become frustratingly difficult to pin down — when pushed, they mention experiments in which researchers have altered the bases of DNA to create a novel, larger molecule that can still store genetic information (H. Liu *et al. Science* **302**, 868–871; 2003). Such fundamental projects are easy to portray as playing God with the stuff of life. But in reality, they have little to do with creating an industry based on new life forms. And even if such work did eventually find applications, most scientists are confident that they could be regulated effectively.

Despite the currently nebulous nature of the 'green goo' scare, the idea may still have resonance with the public. Nanotechnologists would be unwise to dismiss ETC's campaign — particularly as many toxicologists agree that too little is known about the effect on human health of breathing in nanoparticles.

In Britain, the scientific establishment has been relatively quick to respond. In June, the Royal Society and the Royal Academy of

Engineering launched a review of the social and ethical implications of nanotechnology, which should report next spring. In the United States, a bill passed in November recommending some \$4 billion in spending on nanotech over the next four years also calls for a federal advisory panel to look into similar issues. But federal officials remain concerned that US nanotechnologists are poorly equipped to engage in a debate on the risks and benefits posed by their work. "Because of the GM debate, the United Kingdom is ahead of us in developing mechanisms for dialogue," says Julia Moore, an expert on nanotechnology policy at the National Science Foundation in Arlington, Virginia.

It's hard to say whether efforts to address the concerns raised by the ETC group will help to avoid the polarization of views that has afflicted the GM debate in Europe. But if scientists don't get involved, debate about nanotechnology will simply be conducted without them. "I've learnt from the GM debate," says Welland. "It's easy to condemn a technology, but hard to fight back."

Jim Giles

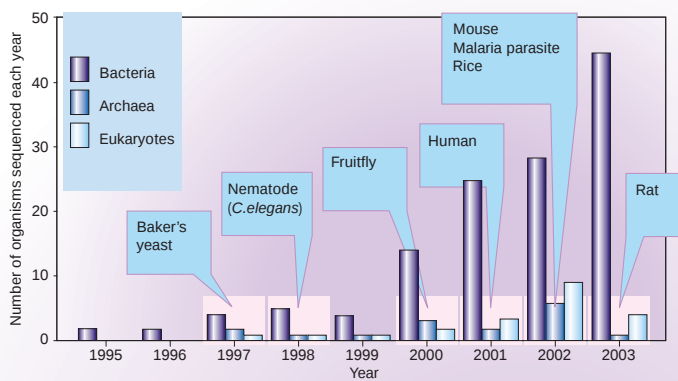
Genomics

Compare and contrast

More, more, more! That is the cry from comparative geneticists, who can't wait to get their hands on newly minted genome sequences. Thanks to the growing availability of such data, their research has gathered unprecedented momentum over the past twelve months. They have gained fresh insights into the regulatory sequences that control the activity of each organism's genes, and into the evolutionary forces that have shaped modern genomes.

"Before, it was like a cottage industry — comparing bits of sequence here and there — but now we're routinely comparing whole genomes with implications that are much farther reaching," says Evan Eichler, a comparative geneticist at Case Western Reserve University in Cleveland, Ohio.

Tracing the 'footprints' of regulatory sequences involves comparing the genomes of several related species. Make multiple comparisons, and conserved regulatory elements should stand out from the 'noise' of sequences shared between close relatives simply through common descent. The more genomes the better, according to researchers who this year completed such comparisons for several related yeast species^{1,2}.



Comparative success: the rising number of sequenced genomes is bringing evolutionary insights.

Analyses of this type are particularly informative if they include both near and far evolutionary relatives — as was illustrated by reports that identified potential regulatory elements in the human genome by comparing chunks of our sequence with those of vertebrates as diverse as other primates, the platypus, chickens and fish^{3–5}.

Comparisons between species have also revised estimates of the total number of genes possessed by important lab organisms. The tally for the baker's yeast *Saccharomyces cerevisiae*, for instance, has been reduced by 500 genes, following the analysis of three related yeast species¹. Meanwhile, the nematode *Caenorhabditis elegans* has gained 1,300 or so genes from comparisons with its cousin *C. briggsae*⁶. Our own gene tally, as recorded on the Ensembl genome-browser website, has slipped this year from about 31,000 to just under 25,000, following comparisons with various other vertebrate sequences.

Comparative analysis has also taught us a thing or two about genome evolution. Over the eons, genomes have been flipped and flopped, added to and chopped, to give rise to new genes and entire gene families. Through cross-species comparisons, it is possible to determine when these changes arose. Comparing various bits of human sequences with those of other primates^{4,7}, for instance, has revealed events that have sculpted our genome throughout evolution. The Y chromosome⁸, for example, seems to have a neat trick for copying important male-specific genes to protect them from being lost as the chromosome degrades down the generations.

Researchers now want to devise experiments to explore the predictions made by comparative genomics. For instance, Mark Johnston of the Washington University School of Medicine in St Louis, Missouri, is putting the proposed regulatory elements identified in yeast² to the test in various ways. One assay involves throwing them onto 'chips' carrying proteins known to bind to regulatory regions of DNA, to see if any of them take the bait.

Carina Dennis

1. Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. *Nature* **423**, 241–254 (2003).
2. Cliften, P. et al. *Science* **301**, 71–76 (2003).
3. Boffelli, D. et al. *Science* **299**, 1391–1394 (2003).
4. Thomas, J. W. et al. *Nature* **424**, 788–793 (2003).
5. Dermizakis, E. T. et al. *Science* **302**, 1033–1035 (2003).
6. Stein, L. D. et al. *PLoS Biol.* **1**, 166–192 (2003).
7. Frazer, K. A. et al. *Genome Res.* **13**, 341–346 (2003).
8. Rozen, S. et al. *Nature* **423**, 873–876 (2003).

www.ensembl.org

Cosmology

Welcome to the real world

It has often been easy for cynics to dismiss cosmologists as whimsical purveyors of 'just-so' stories. But 2003 will go down as the year in which cosmology became a precise observational science — and for that we can thank NASA's Wilkinson Microwave Anisotropy Probe (WMAP), a satellite launched in 2001 that has been studying the faint afterglow of the Big Bang.

In February, WMAP produced an all-sky map, with unprecedented resolution, of the ancient microwaves that fill the Universe. By combining WMAP data with measurements from ground-based microwave telescopes, researchers have pinned down the Universe's age to a relatively sprightly 13.7 billion years,

give or take about 200 million. They have also verified that the cosmos is overwhelmingly made up of two shadowy constituents: dark matter and dark energy. Dark matter, invisible mass whose gravitational pull helps to shape galaxies, seems to make up about a quarter of the Universe, whereas dark energy, which accounts for almost everything else, is a mysterious phenomenon that pushes matter apart at the largest scales. That leaves only about 4% for the visible stuff such as stars, planets and clouds of interstellar gas¹.

WMAP's microwave picture agreed with other results obtained this year, such as measurements of distant, exploding stars². "The degree of cosmic consistency is really heartening," says Charles Bennett, lead scientist on the WMAP project at NASA's Goddard Space Flight Center in Greenbelt, Maryland.

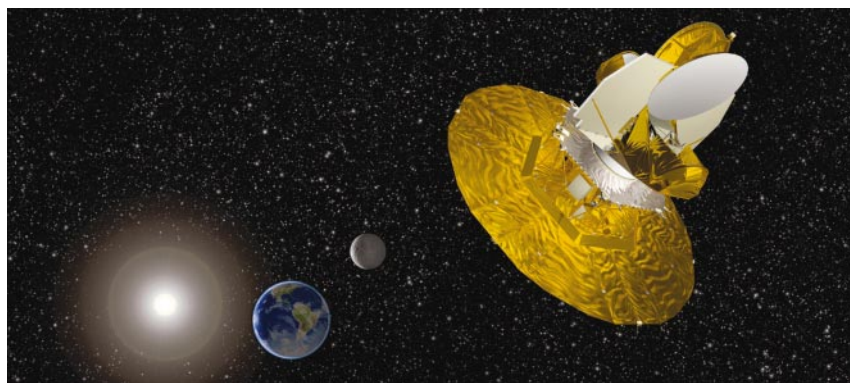
But the data did contain a few surprises. Most notably, fluctuations in the temperature and polarization of the microwaves detected by WMAP showed that the first generation of stars and galaxies appeared just 200 million years after the Big Bang¹ — hundreds of millions of years earlier than most theorists had thought.

Measurements of microwaves from disparate points of the sky also present something of a mystery, according to John Carlstrom, an astronomer at the University of Chicago. The temperature differences between these points match poorly with standard theories, he says. This has led to some new ideas about the shape of the Universe: it could, for example, be dodecahedral, rather like a soccer ball³.

That will give the theorists some thinking to do, as their favourite hypothesis has been that the Universe is flat and infinite. Happily, the shape of the Universe can be probed by future data on the microwave background — from WMAP, ground-based telescopes and the European Space Agency's Planck satellite, which will carry instruments of even greater precision and is set to launch in 2007.

Geoff Brumfiel

1. Bennett, C. L. et al. *Astrophys. J. Suppl.* **148**, 1–27 (2003).
2. Riess, A. G. et al. Preprint at <http://www.arxiv.org/abs/astro-ph/0308185> (2003).
3. Lumet, J.-P. et al. *Nature* **425**, 593–595 (2003).



Glowing report: the WMAP spacecraft has helped to quantify the composition of the Universe.



China

From SARS to the stars

Yang Liwei is a *bona fide* national hero. After clambering from his space capsule after touching down in Inner Mongolia on 16 October, he was hoisted shoulder-high by his ground crew, looking somewhat bemused. Having completed 14 orbits of Earth in a craft based on 1960s-vintage Soviet technology, Yang was lauded by China's state-run media and accorded instant celebrity status. The Beijing Space Medical Institute has even attempted to register his face as a trademark to keep it from appearing on calendars and playing cards.

For the 1.3 billion people of more than 50 ethnicities who inhabit the People's Republic of China, Yang's voyage was proof that their nation is now a major technological power. It is also a leading player in research, according to a report released in October by the Paris-based Organisation for Economic Co-operation and Development (OECD). This placed China third in the global league table of spending on R&D, behind only the United States and Japan, with an annual expenditure of some US\$60 billion. What's more, the OECD's figures dated

from 2001 — since when China's investment in fields from stem-cell research to nanotechnology has continued to blossom.

China's leaders often seem to be obsessed with superlative engineering feats. In July, for instance, the world's biggest hydroelectric project — the Three Gorges Dam in Hubei province — began to generate power. But recent years have also seen a concerted effort to build a solid base in fundamental research. Labs throughout the West have been filled with eager young Chinese students and post-docs, sent abroad for training. And in activities such as field trials of transgenic crops, China is already established as a world leader.

Now China's investment is beginning to bear fruit more widely, with the emergence of research centres and projects that are making waves internationally. This year, for instance, biologists in Shanghai sparked widespread interest with a claim to have generated embryonic stem cells from 'hybrid' cloned embryos derived from human cells and rabbit eggs¹. And in October, China became the first country to license a form of gene therapy for regular clinical use, treating head and neck cancer using the p53 tumour-suppressor gene.

But the dominant story to emerge from China in 2003 was not Yang's joyous homecoming, nor any of the country's various other scientific and technological achievements. Instead, an outbreak of a new respiratory disease, which originated in the rural province of Guangdong, made headlines worldwide. Severe Acute Respiratory Syndrome, or SARS, eventually killed more than 800 people across the globe. Had the virus responsible turned out to be more infectious, the toll could have run into hundreds of thousands. Yet public-

health experts believe that the disease might never have become a global threat had Chinese officials not reacted to the outbreak in such a confused and secretive manner.

For months, China refused to acknowledge that a new virus was circulating among its population — its medical establishment tried to convince the World Health Organization (WHO) that the culprit was the bacterium *Chlamydia pneumoniae*. Officials eventually came clean in March, after the WHO released a global alert about a mystery form of pneumonia that was, by then, turning up across the world. In the months that followed, the SARS outbreak became a symbol of change in China. In an unprecedented move, the minister of health and the mayor of Beijing were sacked for failing to protect the public. "Officials in the future will have to think about whether they are doing their job," says Zihao Rao of Tsinghua University in Beijing, who led a team that determined the structure of an enzyme required for the SARS virus to replicate².

SARS ignited a wave of research activity in China. "The epidemic pushed people," says Nan Shan Zhong of the Guangzhou Institute of Respiratory Diseases, whose team established that the outbreak in Guangdong was indeed caused by the SARS virus³. But some of the resulting activity has been as chaotic as the initial response to the disease. Zhong estimates that more than 100 different research groups are trying, with little coordination, to produce a SARS vaccine. In November, government officials announced plans to begin testing a vaccine based on live, weakened SARS virus, produced by the Beijing firm Sinovac Biotech. But with so many labs working with the virus, some scientists are

HIGHLIGHTS

History on ice

A core some 3.2 kilometres long was pulled from the Antarctic ice, allowing geologists to see at least 750,000 years into the past. Although not the longest ice core ever recovered, it is the oldest, and should contain a record of eight ice ages. Palaeoclimatologists are especially keen to inspect a period some 450,000 years ago when Earth's orbit was very similar to what it is today, giving a good idea of what the present climate might have been like without human influence.

Copy rats

Rats have been a favourite lab animal since the early nineteenth century and, for many studies, provide a better physiological match with humans than mice do. But they missed out on the cloning revolution of the 1990s — the method that gave us cloned mice failed to work. This year researchers came up with a subtly different technique, which involves delaying the activation of the egg, to clone the rodents successfully. Using cloning, it should become easier to produce certain types of genetically engineered rat.

Puzzle solved?

A Russian mathematician claimed to have proved the Poincaré Conjecture, a problem that has gone unsolved since 1904, and which involves properties of objects that do not change when they are stretched or shrunk — such as their number of surfaces. Grigori Perelman of the Steklov Institute of Mathematics in St Petersburg toured American universities describing his solution. Something of an eccentric, he does not plan to submit his work to a peer-reviewed journal.

Flavour saviour

Particle physicists confirmed the mechanism by which the ghostly particles known as neutrinos flicker between different forms, or 'flavours', with an elegant experiment. Japan's KamLAND detector on the island of Honshu picked up neutrinos belted out by a ring of nuclear reactors that coincidentally surround it. Previous experiments have relied on the Sun as a neutrino source. Although physicists have a good idea of the number and type of neutrinos emitted by the Sun, there remained a tiny uncertainty, which the KamLAND experiment neatly eliminated.

Quantum leap

The second of the two basic building blocks of a quantum computer — a 'controlled NOT' logic gate, which flips the state of a target bit of information based on the state of a control bit — was demonstrated by several teams this year. One group managed to make the gate in a solid-state system, using electrons in a silicon chip. The same team also made a solid-state version of the first building block — a 'rotation' gate — in 1999. But no one has yet linked the two gates together to form a functioning computer.



Triumph and tension: a varied year for China saw Yang Liwei (left) feted as a national treasure after becoming the country's first man in space, while on the ground panic spread over the SARS epidemic.

concerned about safety. "We are really worried about the risk of SARS escaping from a laboratory," says parasitologist Zhao-Rong Lun of Zhongshan University in Guangzhou.

Given these problems, wholesale changes are needed before China can truly claim a place among the world's premier scientific, medical and technological powers. The SARS story is symptomatic of a wider failure to tackle public-health issues. Hepatitis B is rampant and AIDS is a growing problem, particularly in Henan province, where tens of thousands of people are thought to have become infected in the 1990s thanks to poor hygiene in a scheme for collecting blood plasma. 'Western' diseases of affluence are also on the rise — obesity among schoolchildren, for example, has risen tenfold over the past eight years.

Meanwhile, China's burgeoning research purse is not matched by a corresponding development of structures to ensure that funds are distributed by merit, or that ongoing projects are properly evaluated. Cronyism is rife, researchers complain. "I constantly have to fight a system that is not used to a merit-based approach," says Muming Pu, who divides his time between the University of California, Berkeley, and his post as director of the Chinese Academy of Science's Institute of Neuroscience in Shanghai.

There is a growing recognition that reform is needed. In September, at a meeting of the China Association of Science and Technology, a semi-governmental body that oversees China's various academic societies, its chair Guangzhao Zhou criticized health officials' unsupported claims about the cause of the SARS outbreak. Presentations at the

meeting highlighted the problems of data-falsification and plagiarism — bold stuff in a country that, until recently, would never have admitted that such misdemeanours take place. Some scientists remain frustrated, however, at the lack of concrete progress towards solving these problems. "Things are not changing fast enough," laments C. L. Tso, a retired molecular biologist who has been pushing for reform for the past 15 years.

Others wonder whether China has its priorities right. Says one molecular biologist at the Shanghai Institutes for Biological Sciences: "Can we really afford a space programme when we have so many people sick and lacking medical care?"

Nevertheless, scientists who have made their names abroad are flooding back to China. Their reasons are many and varied: some are enthused by the chance to run their own research group at a young age; others say they want to contribute to China's development; some are undoubtedly enticed by lavish salaries and research funding.

Botanist DeZhu Li was an early returnee. In 1997 he joined the Kunming Institute of Botany in southwestern China, after three years of postdoctoral work in Britain. He was quickly made deputy director of the institute, which will soon open a new, US\$18-million germplasm bank to store seeds and frozen tissue samples. "It's a really good sign," says Li. "The government cares about science and it has the resources to back this up."

David Cyranoski

1. Chen, Y. *et al.* *Cell Res.* **13**, 251–263 (2003).
2. Yang, H. *et al.* *Proc. Natl Acad. Sci. USA* **100**, 13190–13195 (2003).
3. Zhong, N. S. *et al.* *Lancet* **362**, 1353–1358 (2003).

"Wholesale changes are still needed before China can claim a place among the world's premier scientific powers."

Developing-world health

The fightback starts here

Will 2003 be seen as a turning point in the battle against the infectious diseases — such as malaria, tuberculosis (TB) and AIDS — that blight the lives of millions of people across the developing world? Certainly, big sums of money have now started to flow, and research and control efforts have begun to come together.

This was the year in which US President George Bush pledged to spend \$15 billion over five years to tackle AIDS in Africa and the Caribbean. The Global Fund to Fight AIDS, Tuberculosis and Malaria, backed by donations from the world's wealthy nations, approved projects worth \$623 million in 50 countries, bringing its total spend to almost \$3 billion. And the philanthropic Bill & Melinda Gates Foundation gave out \$700 million for global health — including \$200 million for basic research.

Although most of the new funds are for efforts to control disease, they will have knock-on effects for research, argues Richard Feachem, executive director of the Global Fund. "They represent a major increase in demand, and guaranteed demand at high volumes is what fuels R&D," he says. "The world will know that there is now a buyer for medicines for diseases of poverty." Control funds also pay directly for epidemiological and clinical research, providing substantial new data, particularly on drug resistance, Feachem adds.

AIDS has cruelly exposed the problem of restricted access to expensive drugs in poor countries. But at least new drugs exist for AIDS. For malaria, TB and a host of other diseases, there has been no significant new therapy for decades, because the market-driven drug industry has had little incentive to develop them.

But there are now signs of a new model for drug development for such diseases, in which the public sector works in tandem with the pharmaceutical industry. The Swiss-based multinational Novartis, for instance, has teamed up with the government of Singapore to launch the Novartis Institute of Tropical Diseases, which will focus initially on dengue fever and multidrug-resistant TB. Novartis will provide most of the institute's annual budget of some \$15 million, and will make any drugs it finds available to developing countries without royalties. The centre hopes to punch above its weight by getting governments and philanthropists to pay for the most expensive aspects of drug development, such as clinical trials.

A similar public-private partnership underpins the Drugs for Neglected Diseases Initiative, launched in July, which will seek cures for sleeping sickness, Chagas' disease



Help at hand: malaria sufferers have been given hope by several well-funded disease initiatives.

and leishmaniasis. These efforts join existing partnerships such as the Medicines for Malaria Venture (MMV), the Malaria Vaccine Initiative and the Global Alliance for TB Drug Development. It is too soon to say whether these initiatives will deliver, as their research leads have yet to survive the long trek to the clinic. But all have strong product pipelines, where little or nothing existed previously.

The Gates foundation has spurred the growth of such partnerships, priming them with millions of dollars. In September, it gave the MMV an extra \$40 million to bring to the clinic the four most-advanced drugs in its 21-candidate pipeline.

Even the wildest optimists, however, point out that the money now being spent is a fraction of the sum needed to defeat AIDS, malaria, TB and other devastating diseases. Indeed, it is likely to be decades before we can assess the impact of today's investments in research and control. "They will need to be sustained for more than a generation to make any significant impact," warns Paul Herrling, head of corporate research with Novartis. ■

Declan Butler

Global Fund to Fight AIDS, Tuberculosis and Malaria

www.theglobalfund.org

Bill & Melinda Gates Foundation

www.gatesfoundation.org

Drugs for Neglected Diseases Initiative

www.dndi.org

Medicines for Malaria Venture

www.mmv.org

Malaria Vaccine Initiative

www.malariaivaccine.org

Global Alliance for TB Drug Development

www.tballiance.org

NASA

Trawling through the wreckage

Horror greeted the firework display that lit up the skies above Texas on 1 February. When the space shuttle Columbia disintegrated on its return to Earth, it shocked everyone — even the handful of NASA engineers who had worried whether the craft might have been seriously damaged by an errant piece of foam on take-off. But in hindsight, disaster seemed inevitable. Seven months later, the accident investigation board detailed how complacent NASA had become about the risks of shuttle flight, and how the space agency had long been trying to accomplish too much with too few resources.

The mood was very different from the 'let's-get-back-on-the-horse' reaction to Challenger's loss in 1986. There were no calls to build a replica replacement vehicle. Instead, the investigation board advised NASA to come up with an alternative craft fast, and to phase out the entire shuttle fleet as soon as possible — a decision the agency had been circling for years. Now NASA and its overseers in Congress and the White House are left struggling with how, or even whether, astronaut flights should continue.

The board also called for a national debate on the subject of manned spaceflight. So far, all that has come of this is a series of congressional hearings, with Sherwood Boehlert (Republican, New York), who chairs the House of Representatives' Committee on Science, taking the lead. Boehlert has argued that it is difficult to determine how NASA should proceed until it can be determined what, exactly, its priorities are: should NASA emphasize science, or manned missions of exploration? While confusion reigns, Boehlert helped put a halt to the agency's plan to build a new Orbital Space Plane until it can decide exactly what the vehicle will be used for — other than to serve as a lifeboat for the International Space Station.

As for the space station itself, this was the year in which research on the orbiting laboratory was supposed to be ramped up. Instead, those plans were put on hold, as NASA was forced by the absence of shuttle flights to scale back to a two-astronaut maintenance crew.

The best news for NASA from 2003 is that the year is almost over. The next year looks set to be one of scientific discovery, with a mission to retrieve a sample from a comet, two Mars rovers scheduled to reach their destinations in January, and Cassini due to arrive at Saturn in July. Press reports have

WATCH THIS SPACE

Send in the clones

The year opened with stories of virgin birth. The Raelian cult, which maintains that the first humans were created by aliens, claimed to have produced its own baby clones. Italian fertility specialist Severino Antinori also said he had women volunteers carrying cloned fetuses. But no clone has been brought forth. The United Nations, meanwhile, debated motions to ban human cloning for reproduction and for research. But it was a similar story of build-up without birth — in the end, members agreed to put the decision off for another year.

Physics heavyweights

America's hopes of finding the Higgs boson — a particle that is involved in lending mass to other objects — in the Tevatron collider at Fermilab in Illinois were dashed this summer. Problems with the collider make it unlikely to generate enough collisions between protons and antiprotons to spy the particle. The Large Hadron Collider at CERN, near Geneva, which is scheduled to begin operating by 2007, could offer our best chance of snaring the Higgs.

Water, water everywhere

It was the International Year of Freshwater. But the World Water Forum in Kyoto, Japan, this spring came up with little to solve the world's problems with this critical natural resource. It did not declare access to clean water a human right; neither did it demand that countries negotiate treaties on sharing rivers. The United Nations' goal of halving the proportion of people without access to safe drinking water and sanitation by 2015 seems a remote prospect.

Off the menu

The prospect of Europeans growing and eating large amounts of genetically modified (GM) food remains as remote as ever. Strict European Union laws on the need to label and trace GM crops were passed this summer, which many observers thought would ease the way for the approval of new transgenic crops. But when member states met in December to discuss approving one such crop for sale, the vote remained split. So an unofficial moratorium on new licences stays in place, at least until ministers meet next spring.

Free for all

Will the scientific literature in future be dominated by journals that do not charge their readers? That is the goal of the 'open-access' movement, which argues that the costs of publishing should be borne up front by those who fund research, rather than those who want to read about it. Open-access journals, which charge publication fees, have been proliferating over the past few years. October saw the launch of the most prominent, *Public Library of Science Biology*, which is competing for top biology papers with *Nature*, *Science* and *Cell*.

impulses, are potassium channels that can be switched on and off according to the voltage across a cell's membrane. In May, MacKinnon's group published the first crystal structure of one of these 'voltage-gated' potassium channels^{7,8}. "It's a major breakthrough," comments electrophysiologist Chris Miller of Brandeis University in Waltham, Massachusetts, one of MacKinnon's early mentors.

In the absence of a detailed structure, researchers interested in such channels had used electrophysiological and molecular data to develop a mechanism for voltage gating. Voltage-gated potassium channels were known to consist of four identical protein subunits around a central core, and most researchers had assumed that the voltage-sensitive domains would be situated in the heart of each subunit. They were thought to undergo delicate, screw-like movements in response to changes in voltage, allowing potassium ions to rush through.

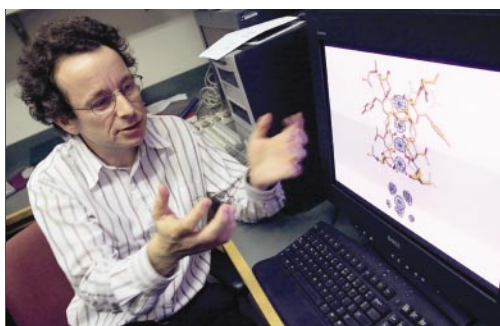
In contrast, MacKinnon's structure places the voltage sensors on the exterior of each subunit, extending into the cell membrane. According to his suggested mechanism, the sensors open the channel by heaving through the membrane in large, paddle-like strokes — more like weightlifting than the molecular ballet envisaged earlier.

Critics claim that MacKinnon's mechanism is unrealistic. "Previous and continuing experimental studies suggest that the paddle model will not hold up," says electrophysiologist Richard Horn of Jefferson Medical College in Philadelphia.

But MacKinnon is undaunted. "I think that these new ideas, which are based on very solid structural and functional data, are correct," he asserts.

Alison Abbott

1. Doyle, D. A. *et al. Science* **280**, 69–77 (1998).
2. Zhou, M., Morais-Cabral, J. H., Mann, S. & MacKinnon, R. *Nature* **411**, 657–661 (2001).
3. Morais-Cabral, J. H., Zhou, Y. & MacKinnon, R. *Nature* **414**, 37–42 (2001).
4. Zhou, Y., Morais-Cabral, J. H., Kaufman, A. & MacKinnon, R. *Nature* **414**, 43–48 (2001).
5. Dutzler, R., Campbell, E. B., Cadene, M., Chait, B. T. & MacKinnon, R. *Nature* **415**, 287–294 (2002).
6. Jiang, Y. *et al. Nature* **417**, 515–522 (2002).
7. Jiang, Y. *et al. Nature* **423**, 33–41 (2003).
8. Jiang, Y., Ruta, V., Chen, J., Lee, A. & MacKinnon, R. *Nature* **423**, 42–48 (2003).



Rod MacKinnon has sparked dispute over his protein studies.

This NASA test shows that insulating foam could have punched a similar hole in Columbia's wing.

also suggested that the White House will unveil a new space initiative — possibly a return to the Moon — within the next few months. Certainly, the agency needs to find itself a convincing new vision, and fast.

Tony Reichhardt

Membrane proteins

Channel voyager makes waves

They are the gatekeepers of the cell, responsible for the movement of ions across its outer membrane. And 2003 was the year in which the scientist who painstakingly deciphered many of their structures received the ultimate honour for his work — just months after his latest data challenged orthodox views in the field.

Beginning with a landmark paper¹ published in 1998, Rod MacKinnon of Rockefeller University in New York has determined the detailed structures of several important ion channels in the cell membrane^{2–6}. Many of his colleagues thought he was embarking on a fool's errand, as proteins embedded in lipid membranes are notoriously difficult to crystallize for X-ray structural analysis. The magnitude of his achievement helps to explain why he was awarded a share of this year's Nobel Prize in Chemistry — a remarkably rapid response from the usually conservative awarding committee.

Indeed, ion channels have not yet yielded all of their secrets, and it was in investigating a remaining mystery that MacKinnon this year ignited a heated debate. Among the most important ion channels, involved in regulating neural

Climate change

The long road from Kyoto

Environmentalists may remember 2003 as the year in which the Kyoto Protocol died. But even if the international community's first attempt at tackling climate change is in terminal decline, this isn't necessarily a defeat for the planet: it might force people to look more realistically at our ability to slow down — and adapt to — the changing climate.

The Kyoto Protocol, agreed in 1997, aims to reduce emissions of carbon dioxide and other greenhouse gases to 5% below 1990 levels by 2012. Although 120 nations are now on board, to come into effect, the treaty must be ratified by a group of countries that together accounted for at least 55% of the world's greenhouse-gas emissions in 1990. The United States, the single biggest emitter, pulled out of the agreement in 2001, leaving only Russia capable of breathing life into the protocol.

As *Nature* went to press, this looked very unlikely to happen. Russian President Vladimir Putin made half-hearted remarks about the treaty at the World Climate Change Conference in Moscow in September. And this month, at a meeting of parties to the Kyoto Protocol in Milan, high-level advisers to the Russian government said that the country would not ratify the protocol, at least in its current form, because it would stifle Russia's economic growth. In the slump that followed the collapse of communism, Russia's greenhouse-gas emissions fell after 1990, giving it a lot of unused 'allowable' emissions that could be sold as 'carbon credits' to other governments. But Russia's emissions, along with its economy, are now on the rise again. Without bankable unused emissions, and without US customers to buy these carbon credits, it looks ever more likely that Russia will say no to the treaty altogether.

"The common view here is that Kyoto doesn't live any more," says economist Henry Jacoby, who co-directs the Massachusetts Institute of Technology's Joint Program on the Science and Policy of Global Change. "It was a heroic effort, but it will simply not work the way it was conceived. It will be necessary to go back, rethink the various elements of the agreement, and start again."

The hope of many is that most of the individual commitments thrashed out during the Kyoto negotiations will live on even if the protocol dies, such as the voluntary commit-



At the extreme: unusual weather conditions saw people struggling against Hurricane Isabel in Virginia (above) and wrestling with bush fires in California.



ments to emissions reductions undertaken by Britain and Germany.

But it remains unclear whether such initiatives will have a significant impact on our climate, which is already feeling the heat. This year, two papers in *Nature* showed that

global warming is having a noticeable effect on living systems. A meta-analysis of studies of more than 1,700 species showed that the geographic ranges of animals are shifting an average of 6.1 kilometres per decade towards the poles¹. And climate change, in connection with other stresses such as habitat destruction, is threatening the survival of a wide range of species, from mammals to trees².

Meanwhile, we have had a taste of the extreme weather conditions that the media have suggested represent what the future will be like under global warming. In Canada, winter rain brought on a deadly avalanche season. The European continent was hit by summer forest fires and a heatwave that, according to some estimates, claimed more than 30,000 lives. Floods hit southern Asia

and China, Hurricane Isabel lashed the United States, and late-season forest fires ravaged the west coast of North America.

It is impossible to say whether these phenomena were a result of global warming. But there is a growing realization that such weather events are becoming more likely. Scientists have found that the patterns of rainfall are changing — European summers are becoming drier, for example, but with more frequent heavy rainfalls and floods³.

How will we face up to the challenges posed by our changing climate in 2004? Earth-system modellers have started to focus on the capacity of smaller regions to adapt to the impacts of climate change. Indeed, regional efforts to combat and adapt to global warming may replace any single overarching scheme such as the Kyoto Protocol. "A homogeneous solution to climate protection is a broken dream," says Hans Joachim Schellnhuber, director of the Tyndall Centre for Climate Change Research in Norwich, UK. "One remaining hope is that it may be superseded by regional initiatives."

Quirin Schiermeier

1. Parmesan, C. & Yohe, G. *Nature* **421**, 37–42 (2003).
2. Root, T. L. et al. *Nature* **421**, 57–60 (2003).
3. Christensen, J. H. & Christensen, O. B. *Nature* **421**, 805–806 (2003).

Neuroscience

Genomics on the brain

The 1990s may have been designated the 'Decade of the Brain', but they will also be remembered as the time when genomics came to the fore. This year, neuroscience and genomics teamed up in projects that promise to propel the study of the brain into the realm of 'big science'.

Working on Parkinson's disease? One of the projects that surfaced this year might provide you with a transgenic mouse that offers fluorescent neurons in key neural pathways. Another could offer you structural brain images along with DNA samples from patients. Still stuck for inspiration? Another project will map all of the active genes in the neural pathways of your choice.

One of the most exciting initiatives owes its existence to the world's fourth-richest man, Paul Allen, who co-founded the software giant Microsoft. In September, he announced the donation of US\$100 million over five years to create the Allen Institute of Brain Science in Seattle, Washington. Its first endeavour will be to produce a comprehensive atlas of gene activity in the mouse brain.

This will be a huge task — two-thirds of the 30,000-or-so genes in the mouse are expressed in its brain. Indeed, the project could generate petabytes of data — the same order of magnitude as all of the information currently held on the Internet. No wonder, then, that the Allen Institute has hired Mark Boguski, one of the world's foremost bioinformaticians, as its founding director.

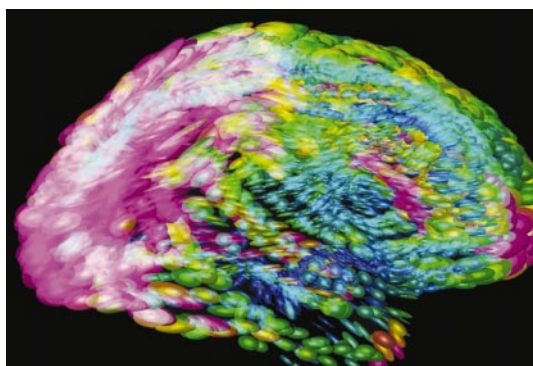
The Allen Brain Atlas picks up from another project with similar goals: the Brain Molecular Anatomy Project (BMAP), launched by the US National Institutes of Health in 1998. "It was too early," says Boguski. "Back then we didn't even know how many genes there were." And whereas BMAP would have taken decades to finish the task, Boguski and his colleagues are refining high-throughput methods that should allow them to complete their atlas by 2006.

The Allen Institute's initiative will show you where all of the genes in the mouse brain are expressed, but if you're interested in gene activity in individual brain cells, your favourite project is likely to be GENSAT — the Gene Expression Nervous System Atlas.

GENSAT's scientists, led by Nat Heintz of Rockefeller University in New York, place each gene under investigation into a bacterial artificial chromosome (BAC), along with a genetic element that stitches a fluorescent tag onto the protein encoded by the gene. By

creating transgenic mice that carry these BACs, the researchers can identify — by sight — the cells in which a gene of interest is normally active, and can work out how these cells connect with the rest of the brain. The same BACs can also be used to introduce other genetic modifications into these cells, providing a unique tool to investigate their biology.

GENSAT's first images can already be seen on its website, which was launched in October alongside a paper introducing the project (S. Gong *et al. Nature* **425**, 917–925; 2003). GENSAT scientists aim to study some 300 genes each year, through funding from the National Institute of Neurological Disorders



All in the mind: a composite image of 20 healthy brains highlighting areas of variation (pink) between the subjects.

and Stroke in Bethesda, Maryland.

This year also saw the launch of the German-led Human Brain Proteome Project, which will catalogue all of the proteins found in the brain. And in the summer, the International Consortium for Brain Mapping, which since 1993 has been compiling a database of structural images of human brains — covering both healthy individuals and patients with neurological or psychiatric disorders — launched itself as a neuroscience resource. Most of the subjects represented in the database have also provided blood samples, which is a boon for researchers who want to investigate the association between particular genetic profiles and unusual brain structures seen in various diseases.

Now that these projects — and related initiatives focusing on particular brain processes or regions — are up and running, perhaps it's time to commission a sequel to the Decade of the Brain. ■

Alison Abbott

Allen Brain Atlas ♦ www.brainatlas.org

BMAP ♦ trans.nih.gov/bmap

GENSAT ♦ www.gensat.org

Human Brain Proteome Project ♦ www.hbpp.org

International Consortium for Brain Mapping

♦ www.loni.ucla.edu/ICBM

GOODBYE

Edward Teller

On 9 September, the father of the hydrogen bomb passed away. He was a controversial figure, who was vilified by many physicists after testifying in 1954 that Robert Oppenheimer, leader of the wartime Manhattan Project, was a security risk. Teller later championed President Ronald Reagan's 'Star Wars' missile-defence initiative. But colleagues at the Lawrence Livermore National Laboratory, co-founded by Teller, say that he should also be remembered for his passion for teaching physics.

Galileo

One of NASA's most successful planetary probes took a suicide dive into Jupiter on 21 September. Having nearly run out of fuel, Galileo was aimed straight at the planet to prevent it from accidentally crash-landing on Europa, one of Jupiter's moons, which some suspect may harbour extraterrestrial life. The craft tore apart as it descended into the gassy planet's atmosphere; its last data transmission arrived on Earth 52 minutes later.

Dolly the sheep

At the relatively tender age of six, Dolly died. After falling ill of a progressive lung disease, the first mammal to be cloned from an adult cell was put down for welfare reasons. Dolly also suffered from arthritis, but it is unclear whether her ailments were a side effect of her unusual genesis. Dolly the icon lives on, stuffed, at the Royal Museum in Edinburgh.

Japan's satellites

Mission controllers were forced to say 'adios' to Japan's Midori-II Earth observing satellite, also known as ADEOS-II, in October, just ten months into a three-year mission. Its electronics are thought to have been blasted by solar flares. The following month, a Japanese rocket carrying two spy satellites was forced to self-destruct ten minutes after launch when its boosters failed to separate. And Nozomi, Japan's first planetary mission, ran out of fuel in December and officially missed its target of Mars — completing a miserable year for the nation's space programme.

And if we're not careful...

Atlantic cod were officially declared endangered by Canada, and stocks have also crashed in the North Sea — where warming waters are reducing larval survival, adding to the effects of overfishing. Cod are not alone. The number of chimpanzees and gorillas in West Africa has plunged by about half over the past 20 years, researchers revealed this year. If current trends continue, hunting, habitat destruction and Ebola fever will drive them to the brink of extinction within a decade. More than 12,000 species of plants and animals face a similarly bleak future, according to the World Conservation Union.

Celebrating 50 years of the cell cycle

To round off a year of scientific commemoration, let's raise a glass to Howard and Pelc.

Sir— This year the world has celebrated the 50th anniversary of the discovery of DNA's structure (see, for example, *Nature* **421**, 395–453; 2003). Meanwhile, however, another important scientific anniversary is in danger of slipping past unmarked.

Also in 1953, Alma Howard and Stephen Pelc published their work on cell proliferation in bean (*Vicia faba* L.) roots¹. They grew plants with a ³²P isotope label and showed that it was incorporated into DNA in the nucleus only during interphase, and that it took 12 hours from the end of division until the beginning of the isotope uptake into new DNA. By analysing heterogeneous populations of meristematic cells, Howard and Pelc deduced that DNA synthesis takes about six hours, and that cells enter prophase of the next mitosis only eight

hours after DNA synthesis is completed.

Howard and Pelc were the first to ascribe a timeframe to cellular life and they proposed the existence of four periods in the cell cycle: a period of cell division, the pre-S-phase (called G1), the S-phase (a period of DNA synthesis) and period G2, or the pre-mitotic period. The concept of the cell cycle was born.

Since then, cell-cycle studies have flourished. It is unfortunate, therefore, that this discovery is now almost forgotten (though not totally: see www.nature.com/celldivision/milestones/full/milestone03.html). The view of the cell cycle formed a basis for determining time parameters of the cell cycle (by labelling mitoses and other methods) and for the biochemical and molecular events that take place at each stage of the life of the cell between divisions in various groups of organisms.

As we know, the concept was later developed and the checkpoints in cell-cycle regulation and universal control mechanisms were determined by using genetics and molecular biology².

All these recent achievements stemmed from Howard and Pelc's study — which calls for another 50-year anniversary celebration to be held by the international scientific community.

Joseph G. Dubrovsky*, **Victor B. Ivanov†**

*Departamento de Biología Molecular de Plantas, Instituto de Biotecnología, Universidad Nacional Autónoma de México, A. P. 510-3, Cuernavaca Morelos 62250, México
†Timiryazev Institute of Plant Physiology, Russian Academy of Sciences, Botanicheskaya 35, Moscow 127276, Russia

1. Howard, A. & Pelc, S. *Heredity* **6** (suppl.), 261–273 (1953).
2. Nurse, P. *Nature* **344**, 503–508 (1990).

What Darwin knew

Sir— In his review of the republication of James Hutton's 1794 book *An Investigation of the Principles of Knowledge and of the Progress of Reason, from Sense to Science and Philosophy*, Paul N. Pearson (*Nature* **425**, 665; 2003) tells us that Hutton devoted an entire chapter to natural selection, and adds, "it seems possible that a half-forgotten concept from his [Darwin's] student days resurfaced afresh in his mind as he struggled to explain the observations of species and varieties compiled from the voyage of the *Beagle*".

Pearson is surely right. But despite his lifelong interest in natural history, Darwin was educated not as a biologist, but as a country vicar. Although he may have read Hutton's book, it is equally likely that Darwin read one of the standard religious works of his day (now perhaps the most ridiculed book in biology), William Paley's *Natural Theology* (1803), which presents Paley's proof of the existence of God, as well as of Divine creation.

Part of chapter five is devoted to what we would recognize as variation and selection. It begins, "There is another answer which has the same effect as the resolving of things into chance." Paley proposes that "the eye, the animal to which it belongs, every plant, indeed every organized body which we can see, are only so many out of the possible varieties and combinations of being which the lapse of infinite ages has brought into existence; that the present world is the relict of that

variety; millions of other bodily forms and other species having perished, being by the defect of their constitution incapable of preservation, or of continuance by generation." As Pearson has commented, Stephen Jay Gould discusses this in his *Structure of Evolutionary Theory* (Harvard Univ. Press, 2002).

Of course, Paley proposes natural selection only to reject it. Nevertheless, it is there. And Darwin himself could not have expressed it better. Natural selection was a heresy in Darwin's day, but a common one.

William L. Abler

Department of Geology,
Field Museum of Natural History,
1400 South Lake Shore Drive,
Chicago, Illinois 60605, USA

There's more to science (and life) than scoops

Sir— Reading Helen Pearson's News Feature "It's a scoop!" (*Nature* **426**, 222–223; 2003) about competition in biology is likely to depress a lot of young researchers. What greater honour or peer recognition could there be than finding competitors relaying the results of your poster by cell phone, as described in your News Feature?

However, I believe they would be wrong to blame journals for accelerated publication or to berate competing labs for beating them to the post. One does not advance science by using faster computers, working in better-equipped labs or even

spending more hours in the lab than could be considered socially normal. Science is best done between your ears. There is no stopping someone who is willing to cut corners to be first in line. The rest of us will just have to think harder and better. Satisfaction comes from knowing you did it right and it was your independent idea, not from a date on the top of your reprint.

When I was leaving the lab of my postdoctoral mentor, Eric J. Brown, I asked him whether he would be pursuing some of the same research that I was about to embark upon independently. He wisely told me that — as we were two different people — even if we did exactly the same experiment one day, we were likely to be performing different experiments the next.

Diversity of thought strengthens the overall progress of scientific inquiry. The formation (and superfunding) of consortia, research institutes and the like brings the risk of stagnation resulting from conformity.

Scoop me once, shame on me. Scoop me twice and I might ask why you aren't capable of any original thought.

I am now going home for the night.

Scott D. Blystone

Department of Cell and Developmental Biology,
SUNY Upstate Medical University,
750 East Adams Street, Syracuse,
New York 13210, USA

correspondence

Contributions to Correspondence may be submitted to corres@nature.com. They should be no longer than 500 words, and ideally shorter. Published contributions are edited.

1904 and all that

This year's anniversaries include neurons, the kodak and enlightenment.

**J. L. Heilbron
and W. F. Bynum**

In 1904 Santiago Ramón y Cajal published the third and last volume of his *Histología del sistema nervioso del hombre y de los vertebrados*, a grand work of 1,800 quarto pages and 887 large engravings. In it, he proved the individuality of neurons, untangled their interconnections with the central nervous system, and set the framework for modern neuroscience.

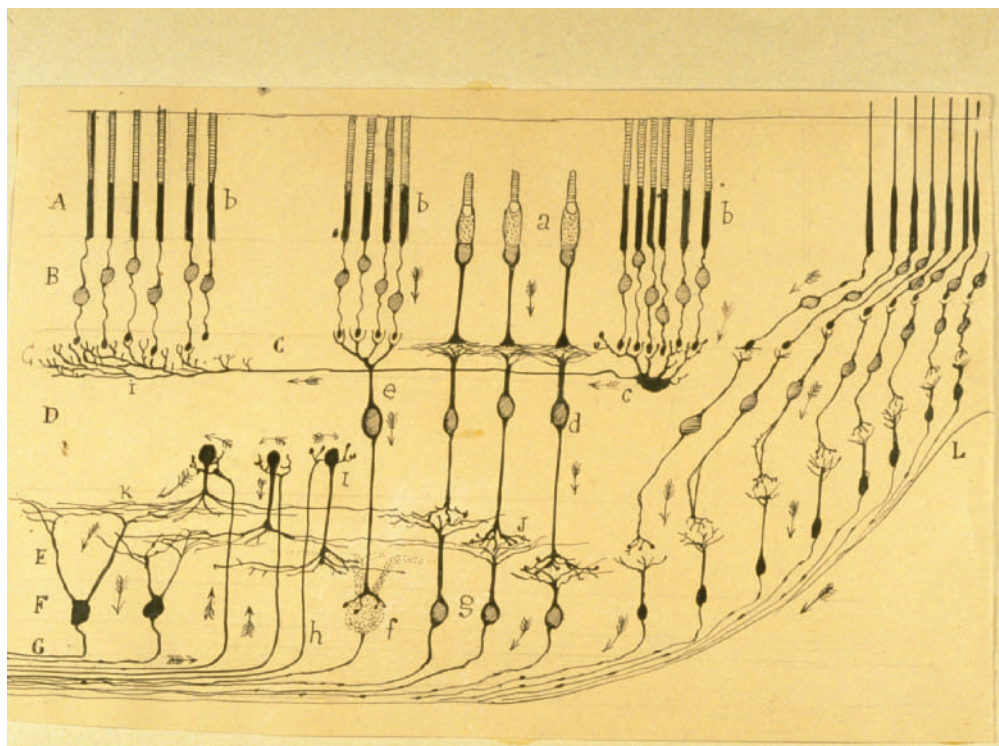
Cajal had put the finishing touches on the first volume as Spain suffered the loss of her last colonial possessions in a humiliating war with the United States. And so he conceived of his work not only as a masterpiece of exact observation and description, but also as a much-needed victory for Spain in international competition. "Above all," he wrote in his autobiography, "I wanted my book to be — please excuse the presumption — a trophy to be laid at the feet of our prostrate national science and an offering of fervent devotion by a Spaniard to his scorned country."

Cajal, who shared the Nobel Prize in Physiology or Medicine of 1906 with Camillo Golgi, the inventor of the staining methods they both used, now has the honour of being our anniversarian of the year.

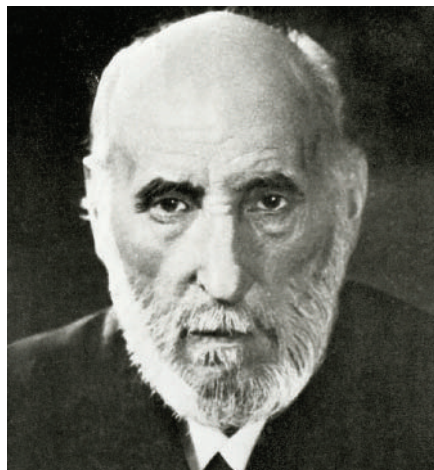
Among the runners-up to Ramón y Cajal's *Histología* were George Boole's *An Investigation into the Laws of Thought* (1854) and Isaac Newton's *Opticks* (1704). Other strong candidates for centennial or semi-centennial celebration are offered below, beginning with 1904 and moving backwards to 1104, when, near enough, Italians figured out how to make brandy. We finish with a quick glance at 1954 and 1954 ± 25 . As usual, ¢ signifies 100 years and \$ a Nobel prize; thus Santiago Ramón y Cajal (\$1906), died 1934 (0.7 ¢).

1904 (1.0 ¢)

We might take exactitude as the theme of 1904. Besides the precise observations and descriptions by Ramón y Cajal, amateurs of the picayune could enjoy the finding by Charles Édouard Guillaume, director of the Bureau International des Poids et Mesures, that a kilogram of water at 4 °C occupied a volume of 1,000.028 cc. That was shocking — the number, by definition, was 1,000.000.



Intricate drawings, such as this observation of the mammalian retinal structure (above), published by Santiago Ramón y Cajal (below) revealed the individuality of neurons and their interconnections.



Guillaume (\$1920, for the invention of invar, an alloy that scarcely expands when heated) did the honourable thing and redefined the litre.

This minor readjustment was not, perhaps, a confirmation of Albert Abraham Michelson's (\$1907) opinion, voiced in 1903, that "our future discoveries must be looked for in the sixth place of decimals". What Michelson had in mind was the discovery of argon, for which Lord Rayleigh and William Ramsay received Nobel prizes in

1904, for physics and chemistry, respectively. The discovery arose from detection of a minute difference in weight between 'nitrogen' taken from the air and nitrogen derived from mineral sources.

In astronomy, Johannes Franz Hartmann was rewarded for painstaking perfection of instruments and exact spectroscopic analysis by the sight of Doppler-shifted calcium lines in light from δ Orionis. This delicate observation, made at the Potsdam Observatory, marked the discovery of interstellar matter, the stuff between the stars and from which they are formed.

On the other hand, we might remember 1904 for work of inspiring inexactitude — the beginning of sustained inquiry into atomic structure. Hantaro Nagaoka, professor of physics at the University of Tokyo, worked out spectroscopic consequences of a Saturn-like atom composed of a large central charge encircled by a great many moving electrons; while Joseph John Thomson (\$1906) tried to understand the periodic properties of the elements by using a model composed of rings of electrons revolving within an 'atomic' spherical space. Nagaoka miscalculated and several characteristics and consequences of Thomson's model could not be sustained. However, the analysis of the

fatal flaws of the imprecise models of 1904 was not fruitless — it proved instrumental in the invention and development of the nuclear atom.

1854 (1.5 c)

In New York City, in 1854, Elisha Graves Otis stood on top of an elevator suspended several stories above the ground and ordered that the supporting rope be cut. A safety mechanism of his invention arrested the descent and made his fortune. Ups and downs and comings and goings make a good theme for almost all our sesquicentennials.

Many ups and downs marked the recurrences of cholera, the most feared disease of the nineteenth-century West. In the middle of Soho, London, John Snow investigated a serious outbreak in the area around Broad Street (now Broadwick Street), and concluded that sewage drains seeping into a public well in the street caused the epidemic. The consequent and famous removal of the pump handle was more symbolic than effective as the epidemic then was already subsiding. Snow's wider epidemiological investigations of the water supply in other parts of London more conclusively demonstrated the water-borne nature of the disease. Also in 1854, the Italian microscopist Filippo Pacini described the culprit, a comma-shaped bacterium, but nobody paid much attention. Robert Koch discovered the same bacterium again in 1884, a wonderful demonstration of the adage that scientific credit often goes to the second person to make the observation or experiment.

Among the newcomers to the world in 1854 were the Catholic University of Ireland, now University College, Dublin; the American inventor George Eastman, who as a bank clerk would learn photography, replace emulsion-covered plates with rolls of coated paper, and invent the Kodak camera; Paul Ehrlich (\$1908), German bacteriologist, who would make haematology into a discipline and discover a not quite magic bullet for the treatment of syphilis; the French mathematician Henri Poincaré, who would impress his colleagues with discoveries across many fields of mathematics, and depress physicists with the news that their profession was on the intellectual level of librarianship; and the Swede Janne Rydberg, the great librarian of spectroscopy, who would sum up the catalogue of the frequencies of spectral lines in comprehensive formulas that would ground the Bohr-Sommerfeld theory of atomic structure.

Among the goers were the Englishman Edward Forbes, Jr, a dredger of molluscs and harvester of starfish, who worked out the biogeography of several lands and seas, and also God's plan in the several creations that brought forth fossils and living forms; the Italian physicist Macedonio Melloni, who spent much of his professional life trying to



Kodak moments: George Eastman (above), revolutionized the way photographs are taken and developed.



Blaise Pascal: established the foundations of probability theory with Pierre Fermat.

discover why some bodies are transparent to light but not to heat, and vice versa; and the German eponym Georg Simon Ohm, who fled an obscure and tedious position teaching mathematics to experiment with electricity and developed the relation between current, electromotive force and resistance that immortalized his name.

1804 (2.0 c)

We propose a double celebration for this double centennial. By mere coincidence, two of the greatest philosophers of the Enlightenment, neither of them French nor otherwise like one another, died in 1804. The elder, born in 1724, was the transcen-

dent idealist Immanuel Kant; the other, the utilitarian realist Joseph Priestley. Kant struggled with the ideas of the great forerunners of enlightened science, Leibniz and Newton, concerning the true natures of space, time, force and matter. He arrived at the doctrine that the human mind contributed space and time to its perceptions of force and matter, which in themselves were unknowable. Priestley did not worry much about how he knew what he knew and increased everyone's knowledge about matter. He was especially prolific in finding new gas types, one of which, later called oxygen, revolutionized chemistry.

As good sons of the Enlightenment, both Kant and Priestley held unorthodox religious views, the one a relaxed deist, the other a combative unitarian. Politically, Kant did not care for the French Revolution; Priestley embraced it, as he had the American. Consequently Kant lived peacefully in the university town of Königsberg and Priestley was driven out of Birmingham by a mob that did not like his political or religious views. Neither man welcomed the main immediate consequences of the new thoughts he unleashed. Kant opposed the radical romantics who claimed to be following him in building a nature philosophy from introspection, and Priestley repudiated the new French chemistry of Antoine Lavoisier.

1754 (2.5 c)

Still in the Enlightenment, the year 1754 saw the publication of one of its most important and popular tracts on the working of the mind, the *Traité des Sensations* by Etienne Bonnot, abbé de Condillac (born 1714). Here he introduced the not altogether original image of a statue gradually endowed with senses and sense experience. The enlivened

statue falls short of full humanity, however, by not being expert in algebra, the most perfect and clearest of languages, according to Condillac's *Logique* (1780). This analysis, like Priestley's pneumatic discoveries, had a spectacular pay-off in the chemistry of Lavoisier, who invoked Condillac on the need for a new chemical nomenclature.

Another gas, carbon dioxide, began its scientific identity 250 years ago in Edinburgh, when Joseph Black's medical thesis identified a peculiar 'fixed air' as a product of the heating of magnesias alba. And another of the world's great sources of hot air (and more solid matter), King's College, later renamed Columbia University, was founded in New York City.

1704 (3.0 c)

The world cannot have too many celebrations of the Enlightenment. This year we can observe the tercentenary of the death of one of its major inspirations, John Locke, and of the publication of two of its key texts. One of these, the *Lexicon technicum* of the London clergyman John Harris, was the first general scientific encyclopedia to draw on the work of Newton and Boyle. It became a primary source for Ephraim Chambers' famous *Cyclopaedia, or an universal dictionary of art and sciences* (1728), which, in turn, inspired Denis Diderot and his publisher to undertake the *Encyclopédie*. The second key text, Newton's *Opticks: or, a treatise of the reflexions, refractions, inflexions and colours of light*, became in its successive editions the exemplar of physical investigation and penetrating insight. Newton's play with prisms showed how sustained experimentation could lead to grand, novel and reliable theory in a limited domain of nature, and its appended "queries" offered broad hints to explorers of many other domains.

Anton Maria Valsalva did not need Newton's hints. Educated in a Jesuit college and at the University of Bologna under Marcello Malpighi, Valsalva excelled as a surgeon and anatomist. His masterwork, *De aure humana tractatus* (1704), provided an exact description and physiology of the ear, some parts of which Valsalva distinguished for the first time.

1654 (3.5 c)

Although Jakob Bernoulli, the founder of the famous family of Swiss mathematicians was born in 1654, the year, mathematically speaking, belonged to Blaise Pascal. He completed a profound book, never published, on conic sections and projective geometry, printed his *Traité du triangle arithmétique* and, in correspondence with Pierre Fermat, established the foundations of probability theory. Not much further progress occurred in the calculus of probabilities, however, for half a century, until Jakob Bernoulli and others turned their attention to it.

1604 (4.0 c)

Politically correct thinkers might be inclined to give King James I a smokeless quadricentennial party for his *Counterblaste to tobacco* had he not earned their unshakeable condemnation for his earlier *Dæmonologie* (1597), a sprightly exhortation to all and sundry to root out witchcraft. Others will prefer to honour that son-of-a-witch Johannes Kepler. While James fulminated against tobacco, John's mother did her sorcery and he published the fundamental principles of geometrical optics and their applications to astronomy in his *Ad Vitellionem paralipomena (Astronomiae pars optica)*. The exercise prepared Kepler to

explain the operation of the telescope to Galileo, who might also be celebrated this year for having written down 400 years ago the law of free fall, though without a sound demonstration.

1504 (5.0 c)

It was half a millennium ago, on leap day 1504, that Christopher Columbus (it is said) frightened the natives of Jamaica by calling on God to darken the Moon. Columbus relied for the threat on the *Ephemerides* published by Regiomontanus in 1474. According to the story, the terrified natives agreed to do the bidding of the man who could command the Moon.



Doctor Phlogister: Joseph Priestley was ridiculed for his political and religious views.

HUTTON ARCHIVE

1204 (8.0 €)

Eight hundred years ago Rabbi Moses Ben Maimon (Maimonides) died in Egypt, where he had been active for many years as court physician and head of the Jewish community. His easy-going rationalism, which made room for science in religion, for philosophy in theology, and for morality in government, influenced Western thinkers from Thomas Aquinas to Leibniz and Spinoza as well as Arab and Jewish scholars in Spain and the Maghreb. He should be resurrected, although he did not believe in it.

1104 (9.0 €)

We need not explain why the invention of the distillation of alcohol from wine, accomplished in Italy in or around 1104, justifies merry making.

1954 (0.5 €)

We begin with the positive. CERN, the European particle-physics laboratory near Geneva, celebrates its fiftieth anniversary this year. Despite the many pressures within it and tensions from without, it still stands as a monument to the spirit of cooperation in Europe and, with its big accelerators, remains at the forefront of particle physics. We trust that the fulfilment of the physicists' dreams of a final theory will not also end the career of the world's premier international laboratory.

Still with the positive, Harold Hopkins and N. S. Kapany described, in this journal, their flexible fibrescope which, appropriately modified, became a mainstay in medical diagnosis and treatment, and John Enders (\$1954, for work on the poliomyelitis virus) and Thomas Peebles isolated the measles virus.

On the negative side, cold war concerns launched the first US nuclear submarine, the *Nautilus*; deprived J. Robert Oppenheimer of his security clearance after a bruising public hearing that exposed the practices of the security state; and poisoned Japanese fishermen with radioactive fallout from the tests of a thermonuclear 'device' at Bikini. An American congressman added insult to their injury — from which one of them died — by suggesting that the fishing boat, the *Lucky Dragon*, was a Soviet spy ship.

One of those energized to campaign against atmospheric testing by the incident of the unlucky dragon was Linus Pauling (died 1994), the winner of the Nobel Prize in Chemistry in 1954 for his work on the chemical bond. Nine years later he received the Nobel Prize for Peace.

Further bad news, Enrico Fermi (\$1938), Italian-American physicist, author of beautiful theories of elementary particles and inventor of the nuclear reactor for plutonium production, died; as did Fritz London, German-American physicist, who with his brother Heinz created the major equations describing superconductivity; Auguste Lumière, who with his brother Louis invented



Rocket science: a still from the film *Woman in the Moon*, on which Hermann Oberth worked as consultant.



the first successful movie camera and the film newsreel; and Alan Turing, the code breaker and computer theorist, by his own hand, after confessing to his homosexuality.

We do not know whether to regard the introduction of TV dinners in the United States as good news or bad.

1954 ± 25

In 1929 Robert Goddard, the grandfather of American rocketry, launched the first instrumented rocket (it carried a thermometer, barometer and small camera) and Hermann Oberth (born 1894), Goddard's German counterpart, designed a manned rocket and spacecraft for Fritz Lang's film *Woman in the Moon*. In 1979 the expectations of these pioneers were realized by Voyager I's flyby of Jupiter (which detected a ring around the planet); they would not have been surprised, however, by the fate of the abandoned Skylab, blown to a fiery death in Earth's atmosphere by a particularly strong solar wind.

We might also compare the electroencephalograph or brain-wave detector of

the German psychiatrist Hans Berger, invented in 1929, with computerized axial X-ray tomography or CAT scanning, for which the South African-American physicist Alan Cormack (born 1924) and the English electrical engineer Godfrey Hounsfield shared the Nobel Prize for Physiology or Medicine for 1979.

We can also celebrate this year the unexpected union of the incomprehensibly large with the unspeakably small. In 1929 Edwin Hubble

announced his famous law relating distances of galaxies to the speed at which they are receding (about 100 miles a second for every million light years). It agreed well with the relativistic expanding Universe. While racing from the Big Bang, the mixture of matter and energy that would make stars and other things quickly cooled to the point that the electroweak force froze into two distinct forces that human physicists have discovered only relatively lately. Three of these — Stephen Weinberg, Sheldon Glashow and Abdus Salam — shared the Nobel Prize in Physics in 1979 for perceiving the form of original electroweakness. At the same time, experimenters at Deutsches Elektronen-Synchrotron in Hamburg revealed something about the strong force by detecting its field quanta, the paste of nuclei, which ever-playful physicists call 'gluons'.

J. L. Heilbron is at Worcester College, Oxford OX1 2HB, UK. W. F. Bynum is at the Wellcome Trust Centre for the History of Medicine, University College London, 24 Eversholt Street, London NW1 1AD, UK.

Getting it off the ground

Celebrating the centenary of controlled, powered flight.

The Wright Brothers and the Invention of the Aerial Age
by Tom Crouch & Peter L. Jakab,
National Geographic Books: 2003. 256 pp.
US\$35, Can\$55

Wings of Madness: Alberto Santos-Dumont and the Invention of Flight

by Paul Hoffman
Fourth Estate: 2003. 384 pp. £10.99
Hyperion: 2003. 369 pp. \$24.95

Andrew Nahum

A hundred years ago this week, Orville and Wilbur Wright made the first controlled, powered flight in a heavier-than air machine. By then it was becoming accepted that the idea of controlled flight might not be so fantastical after all. The scientific analysis of air and flying machines had begun to be taken seriously in the nineteenth century by engineers and scientists such as Gustave Eiffel, Samuel Pierpont Langley and Hiram Maxim.

With the exception of Otto Lilienthal, who flew his hang-gliders repeatedly, few realized how capricious the air is. Although there was much debate about the correct expression for the resistance of the air, which would determine wing lift, the more immediate issues of pitch stability (controlling movement up and down), balance and control determined how long a 'flight' lasted before the almost inevitable crash. How many seconds this took seems rather unimportant now, but it still receives the attention of many partisan enthusiasts determined to prove that Richard Pearse in New Zealand, Langley in Washington, the German-American Gustavus Whitehead, or Clement Ader in France flew, or even 'could have flown', before the Wright brothers.

The reason for the Wright brothers' success was that they combined in their joint enterprise the best empirical skills of Yankee inventiveness, intelligent observation and a scientific, or at least systematic, investigation into flight. Furthermore, as cyclists and bicycle-makers, they understood the importance of piloting.

They sensed, in a way that almost no other flight pioneer except Lilienthal and the engineering scientist Percy Pilcher did, that a successful flying machine would not spring from the drawing board like a steam ship, ready to plough the aerial ocean. They knew that a flying machine might be successful but would still be nervy and treacherous, subject to the air currents of the moment. A bicycle might be an excellent machine, but the learner might still fall. They appreciated

that learning to fly was different to the art of creating a successful machine, and set out to do both, or, in Wilbur's words, "to escape accident long enough to know how to avoid accident".

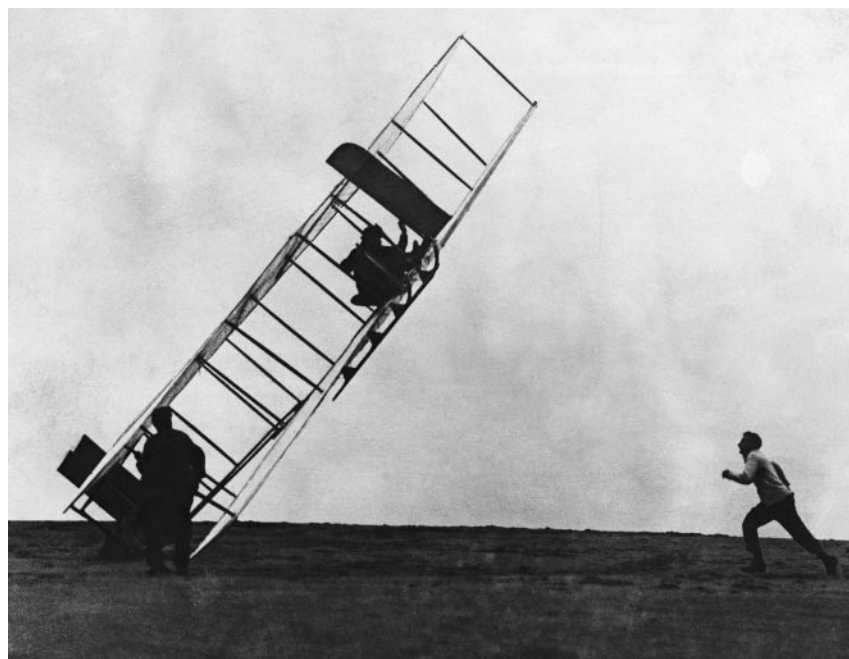
It is impossible to know exactly what each of the Wright brothers contributed to the venture, so close were they in everything. But of the two, Wilbur, who was destined for Yale before suffering some strange neuroathletic breakdown, seems to have been the more analytical. It was Wilbur who wrote of observing the flight of buzzards, and that "they regain their lateral balance, when partly overturned by a gust of wind, by a torsion of the tips of their wings". This observation translated into what was perhaps the Wrights' most notable achievement: the realization of the importance of control in the roll axis. Their contemporaries, in contrast, thought that adequate control and steering could be achieved purely by an elevator (pitch control) and a rudder (yaw control). The Wrights, however, understood that successful turns actually require the aircraft to bank, using a proportion of the lift force to initiate and sustain the turn.

Wilbur's systematic nature also led him to enquire about average wind speeds around the United States. He reasoned that because "it is necessary to move through the air at 15 or 20 miles an hour in order to obtain support, it is safer to practice in a wind, provided this is not too much broken up into eddies

and sudden gusts by hills, trees and so forth". This insightful statement shows that he visualized the wind as a smooth, ideal flow, until it was perturbed by obstacles. This meteorological requirement took the brothers to Kitty Hawk, on an outer bank of the North Carolina coast where, during experimental sessions in 1900, 1901 and 1902, they developed their wing-warping gliders until they could fly smoothly and consistently.

To design the wing they also collected and collated data from Lilienthal and from their own wind-tunnel experiments on wing camber, coefficients of lift and drag, and the movement of centre of pressure with changing incidence. Their analysis indicates considerable mathematical ability for boys who left school at 17, and is a remarkable tribute to the American public education system. At the same time they developed their piloting skills and their understanding of aircraft control so that, in the 1902 season, they were able to complete more than 700 successful glides, returning to Dayton to build an engine and attempt controlled, powered flight. None of their contemporaries made such persistent progress to airmanship or developed full, balanced, three-axis control (of pitch, yaw and roll) until the Wrights finally demonstrated this publicly in 1908.

The excellent book by Tom Crouch and Peter Jakab is one of the first to speculate on the brothers' mentality, their home life and



Trial and error: not all of the Wright brothers' test flights using gliders were entirely successful.

HULTON-DEUTSCH/CORBIS

the source of their originality. It contains a generous ration of the wonderful photographs taken by the Wright brothers of their gliding experiments, and forms an engaging account of these epochal events. However, those wanting more depth must go to the two authors' earlier books: Tom Crouch's *The Bishop's Boys* (W. W. Norton, 1989) and Peter Jakab's *Visions of a Flying Machine* (Smithsonian Institution Press, 1990).

The aviator and celebrity Alberto Santos-Dumont continues to puzzle and intrigue, but many details of his life have been obscure, his own book *My Airships* being one of the few sources of information. So Paul Hoffman's *Wings of Madness* is a much-needed and long-overdue account. It concentrates more on Santos Dumont's life than on the details of his flying machines. This fabulously rich Brazilian coffee tycoon settled in Paris and took his first free balloon flight with a professional balloon-maker in 1897, proving his scientific credentials by ascribing the extra effervescence of the champagne he took aloft to the high altitude. He soon progressed to small, light airships, initially using the remarkably light internal combustion engine developed for the De Dion-Bouton tricycle. Santos-Dumont became a popular figure as he dashed and sometimes crashed around Paris, often demonstrating remarkable control and mooring the craft outside his apartment on the Champs-Élysées.

By 1906, Santos-Dumont, stimulated by the move among French aeronauts to better what they had heard of the Wrights' flight, had built a heavier-than-air craft. The diminutive Santos-Dumont stood in the vehicle as it made a wallowing 240-metre flight in the Parc de Bagatelle. This public flight attracted media attention but the craft was barely under control and made no contribution to heavier-than-air flight.

When Wilbur Wright came to France in 1908, Santos-Dumont seemed put out by the adulation that the American received, and began to show the beginnings of the depression and mental problems that would subsequently dog him. He became obsessed with new uses for aeroplanes. The final straw,

which caused him to take his own life, was seeing the use of aeroplanes in the Brazilian revolution of 1932. His last recorded words are said to have been: "I never thought my invention would cause bloodshed between brothers. What have I done?" ■

Andrew Nahum is senior curator of the Aeronautical Collection at the Science Museum, Exhibition Road, London SW7 2UD, UK.

A healthy draught of scepticism

Eight Preposterous Propositions: From the Genetics of Homosexuality to the Benefits of Global Warming
by Robert Ehrlich

Princeton University Press: 2003. 360 pp.
\$27.95, £18.95

Walter Gratzer

A little way into Robert Ehrlich's assault on obfuscation and unreason, your appalled eye will light on a table which reveals that more than a quarter of the population of the United States believes in witches, 41% in possession by the devil, fully a half in extrasensory perception (ESP), and no less than 45% are in no doubt that extraterrestrial beings have been stalking the Earth. (The physicist Leo Szilard said so too, but added that they are called Hungarians.) Worse still, even among the beneficiaries of a college education, only 16.5% are prepared to concede that *Homo sapiens* is a product of evolution, unaided by the hand of God.

Such dense fog between the ears is invariably linked to an inability to grasp that improbable events are merely manifestations of the rules of chance, and not of divine intervention. Oscar Wilde understood this ingrained disorder of the human intellect: "Man can always believe the impossible, but man can never believe the improbable," he observed. Ehrlich has set himself the heroic task, concealed beneath

his flippant title, of confronting the tide of irrationality in what is in effect a manual of scientific reasoning.

His method, originating in his earlier book *Nine Crazy Ideas in Science*, is to test eight quite diverse propositions, extending from the unquestionably absurd (telekinesis, or moving matter around by thought alone) to the probably valid, such as a part for genetic factors in determining sexual inclination. He grades these on a scale of 'flakiness': zero flakes implies that the proposition may well be true, and four flakes that it is unarguably nonsense. My dictionary defines 'flaky' as "adj. eccentric, crazy", but this is not altogether what Ehrlich means by it; he conceives it as "lacking in empirical evidence or internal consistency", thereby distinguishing it from his 'crazy ideas' in science, some of which (like practically all of Kuhn's 'paradigm shifts') could be true.

Ehrlich's longest chapter is devoted to the weighty question: "Should you worry about global warming?" After an impeccably neutral analysis of the passionate opinions on either side, Ehrlich awards it a (judiciously qualified) 'one-flake' rating — in other words, a negative answer to the question could just about be entertained. A strength of Ehrlich's treatment is that he approaches the truly preposterous theses — telekinesis and the eclipse of evolution by "intelligent design" — with a straight face. His reasoned demolition of the evidence for these aberrations is vastly more effective than the red-eyed apoplexy that seizes the average scientist at their mere mention.

But it is the final two chapters that I found the most compelling. Ehrlich is at his most incisive on the placebo effect, and on the recent assertion in a widely publicized paper that its extent has been grossly exaggerated. The arguments hinge, for the most part, on the interpretation of statistics, which Ehrlich manages to make accessible to all who will make the effort. Only at one point, isolated in a box from which innumerate readers can avert their eyes, does he set out the mathematics in full.

He makes a powerful case that many, and especially psychotropic, drugs which make extravagant profits for the pharmaceutical industry are ineffective or worse. He uncovers the weaknesses in conventionally designed double-blind trials and, both here and in his final chapter ("Should you worry about your cholesterol?"), he expatiates on the lax standards by which the industry is now regulated, and the way in which the once-proud US Food and Drug Administration has been emasculated. What is especially striking is the low confidence limit ($P < 0.05$) considered adequate to establish the efficacy of a new drug in clinical trials — a level that is perhaps acceptable in sociological research, but is generally considered risible in the exact sciences.

Let it snow

It is said that no two snowflakes are identical. Each contains about a billion billion water molecules, so the number of possible configurations is enormous. These photographs by Patricia Rasmussen are from *The Snowflake: Winter's Secret Beauty* by Kenneth Libbrecht (Voyageur Press, \$20).



Ehrlich records the concern of one psychiatrist that if antidepressant drugs were shown to be ineffective, patients would suffer by being deprived of the benefits of the placebo effect. A striking study, published last year, revealed that in PET scans the same region in the brains of patients became active whether they were given placebos or opioid-receptor drugs. What we clearly need, then, are better placebos.

There is a short story by the US humorist Josh Billings in which a farmer discovers that his black horses are eating more hay than his white horses. Eventually he gets to the root of the matter and finds the explanation: he has more black horses than white. The confusion of mind that Ehrlich exposes to view is often close to that of Billings's farmer.

Here and there I found myself wondering whether, in striving for objectivity, Ehrlich was not putting too much weight on publications of uncertain authority in dubious journals. All the same, he has dug consistently deep and marshalled the evidence in masterly style. He is unfailingly lucid, and if his colloquial, rather jokey manner brings his book more readers, then so much the better, for the lessons that it teaches are important to us all. It upholds, moreover, the principle enunciated by an academic politician of legendary dexterity, a famous British vice-chancellor: one must always keep scientists away from committees — they are apt to change their minds in response to the evidence. ■

Walter Gratzer is emeritus professor of biophysical chemistry, King's College London, the Randall Centre, Guy's Campus, St Thomas' Street, London SE1 1UL, UK

Birds and the double elephant

Audubon's Elephant: The Story of John James Audubon's Epic Struggle to Publish *The Birds of America*

by Duff Hart-Davis

Weidenfeld & Nicolson: 2003. 192 pp. £18.99 (to be published in the US by Henry Holt next April)

Audubon in Edinburgh: And his Scottish Associates

by John Chalmers

National Museums of Scotland: 2003. 240 pp. £30

David Knight

John James Audubon's enormous *Birds of America* books (1827–38) could not be printed and published in his native United States — the skills available to produce a book of this size were available only in Europe. So Audubon travelled to Britain,



Huge undertaking: this yellow-billed cuckoo was life-sized in John James Audubon's *Birds of America*.

first to Liverpool and then on to Edinburgh, bringing with him his astonishing portfolio, which was big enough for all of the birds to be portrayed life-sized. *Audubon in Edinburgh* and *Audubon's Elephant* both focus on how the resulting extraordinary volumes of *Birds of America*, which now sell for millions of dollars, were produced. And they tell how Audubon recruited subscribers to fund the printing of successive volumes, a process that took more than a decade. It is a romantic story.

Audubon's volumes stand a metre tall — the title of Hart-Davis's book refers to the size of the format, double elephant — and his work is marvellous, but is it science? Or is it, in fact, a white elephant? This question in part lay behind Audubon's rows with George Ord of the American Philosophical Society in Philadelphia and the eccentric traveller, squire and naturalist Charles Waterton in England, who considered him a charlatan. Clearly, Audubon was not that, even if his tales from the wild frontier about wildlife, ancestry and adventures improved in the telling. In line with the maxim "What's hit is history, but what's missed is mystery", Audubon slaughtered masses of his beloved birds in order to mount them convincingly and paint them.

Huge tomes sold in small editions are not the stuff of serious ornithology. This made Audubon's alliance with the Scottish naturalist William MacGillivray so important, although in print Audubon hardly seems to have done him justice. MacGillivray did the dissections and got the science right in the volumes of text, *Ornithological Biography*, that were published to accompany the plates. William Swainson had been in line for this

job but had stuck out for fairer terms and broken off negotiations. Swainson was a leading illustrator and author of books on zoology, but his trinitarian (or quaternian) taxonomic system was coming to be seen as a speculative straitjacket on nature, and the book would have been poorer had it been saddled with it.

Audubon's life until 1826 had been a series of failures culminating in rows in Philadelphia where the memory of Alexander Wilson, his predecessor as America's leading ornithologist, was assiduously kept fresh. But despite his detractors, Audubon was elected to several scientific societies, including the Royal Society. The books by Hart-Davis and Chalmers both tell the tale of how this frontiersman, for whom English was a second language, managed to break into the scientific community through his energy, enthusiasm and sheer talent as a painter.

Finding friends and admirers in Liverpool, Audubon went to Edinburgh with more confidence. There he ignored advice to publish in a small format, and met William Lizars, who was to do some exquisite plates for William Jardine's series *The Naturalist's Library*. Lizars engraved and printed Audubon's plates and coloured them by hand — this was by necessity variable, and was a frequent source of complaint from the subscribers who sponsored the work. After producing a few of these, Lizars, faced with a strike and overstretched, was happy to give up.

Audubon, by then in London, found that Robert Havell and his son (also Robert) would do the job at a lower rate, and they ended up doing almost all of the plates.

Audubon's was one of the last great books to be illustrated with engravings, as Swainson and others perfected the cheaper method of lithography. Each volume generally contained, unsystematically organized, some plates of one or two big birds, and some of several smaller ones. But Audubon struggled to maintain enough subscribers, as some died or fell away as the work progressed. Even so, he triumphed, even in the United States, to which he returned several times.

Audubon's Elephant and *Audubon in Edinburgh* both recount his eventful history.

But whereas Hart-Davis has produced a compact, elegant and readable biography, Chalmers' less-organized volume is focused as much on Edinburgh and its scientific world as on Audubon.

The size of Chalmers' book works better with Audubon's plates — the bigger the better, as he intended. And whereas Hart-Davis reprints only Audubon's pictures, Chalmers has also included many by other contemporary artists, so that we can compare their work and get a real feeling for Audubon's achievement. This also reveals the demands placed on anyone depicting nature

at a time when Charles Darwin was a student.

Although most of the plates were printed in London, Audubon hated the bustling, foggy metropolis, but loved Edinburgh. Chalmers puts across a vision of that city and its medical institutions when Wordsworth's line "We murder to dissect" was being put into practice by Burke and Hare. The book places Audubon securely within the science of his time.

David Knight has taught the history of science at the University of Durham, UK, since 1964.

His latest book is Science and Spirituality: The Volatile Connection (Routledge, 2003).

Science and culture

Nasal bioluminescence in reindeer, *Rangifer tarandus rubens*

Martin Kemp

Centre for Visual Studies, University of Oxford, Oxford OX1 2PT, UK

The discovery in 1939 of a reindeer mutant characterized by a luminescent snout (and popularly known as 'red-nosed') is a neglected chapter in the history of selective breeding. The mutation initially resulted in the young fawn (nicknamed Rudolph) being ostracized, in much the same manner as documented by H. Christian Andersen in his classic account of the so-called 'ugly duckling'¹. However, the timely intervention of S. Claus, who breeds reindeer in Russia's Chukotka peninsula, ensured the persistence of the new trait, which

has proved navigationally beneficial on a world-wide basis². Here I show that the precise mechanism for the luminescence in the reindeer *Rangifer tarandus rubens* is identical to that which causes the glow in railroad worms.

There has been unwarranted confusion about the first record of the red-nosed reindeer. Priority belongs to R. L. May of the Montgomery Ward department store in Chicago, who made the first independent sighting, accompanied by draftsman D. Gillen, who clearly recorded its shining nose in his sketch and printed illustration³ (Fig. 1). The

resulting publication was given the widest possible distribution, to such effect that six million copies had been sold by 1946. The fame of the mutant was ensured by the popular, if inaccurate, musical rendition of its discovery by G. Autrey in 1949. It is now known that selection for the Rudolph mutant by the notoriously secretive Chukotkan herder, S. Claus, actually about came by chance, during a journey across the habitat occupied by the young reindeer⁴. Claus recognized that nasal bioluminescence presented a significant adaptive advantage in the prevailing meteorological conditions. The records for 25 December 1949 indicate a worldwide fog of unprecedented extent⁵. Claus' adoption of Rudolph and cultivation of a pure-breeding strain facilitated his navigational tasks during the smogs that beset developed countries in the 1950s and 1960s.

The famous red glow arises from a chemiluminescent reaction involving a direct conversion of chemical energy into light. It involves the substrate D-luciferin combining with ATP and oxygen, in a reaction catalysed by the enzyme luciferase, with light being given off along with the product oxyluciferin. The forms of luciferin and luciferase differ chemically in different organisms⁶, with the colour of light given off varying according to the substrate, ranging from blue in marine organisms to yellow in fireflies. The mechanism in *R. t. rubens* is identical to that seen in larva of the beetle *Phrixothrix hiatus*, known as the railroad worm, which produces red light from its head and green light from its body⁷.

The size of the current population of Rudolphs is impossible to estimate with any accuracy, but it must run into many millions. There have been proposals to classify the twentieth-century 'red-nosed' subspecies as a pest⁸. However, a proper understanding of its natural history will almost certainly reveal that natural controls are already beginning to operate effectively.

Received 27 November 2003; accepted 6 December 2003.

1. Andersen, H. C. The "Ugly Duckling" Phenomenon in Duck and Swan Populations. *Danish Biol. Rep. Int. Ser.* **944**, 199–205 (1953).
2. Christmas, Fr *Memoirs of S. Claus, Reindeer Herder* 169 and 222–256 (Almqvist & Larsen, Stockholm, 1950).
3. May, R. L. *Rudolph the Red-Nosed Reindeer* (Maxton, New York, 1939). Also letter of 24 December 1939 from Denver Gillen to Daisy Gillen, in *Gillen Papers* 639/24, University of Lucerne, Texas.
4. Claus, S. *Memoirs* 224–227 (1944).
5. *Am. Meteorol. Surv.* **91**, 561 (1950).
6. Luce, M. & Seurat, P. in *Bioluminescence in Marine and Terrestrial Organisms. A Comparative Study* (transl. Signac, P.; Flammarion, Paris and New York, 2000).
7. Treno, F. S. Rosy luminescence in the heads of railroad worms. *J. Brazil. Biol.* **69**, 189–194 (2001).
8. Scrooge, E. A population out of control. *Int. J. Hunting* **590**, 15–440 (2002).

Acknowledgements I thank A. Abbott for her work on catalysis, which she has shared with me. C. Dawney provided vital support in helping me obtain PFD funding for this research.

Competing interests statement The author denies that he has any intention to obtain any financial interest in Rudolph.

Correspondence and request for materials should be addressed to the author (martin.kemp@trinity.ox.ac.uk).

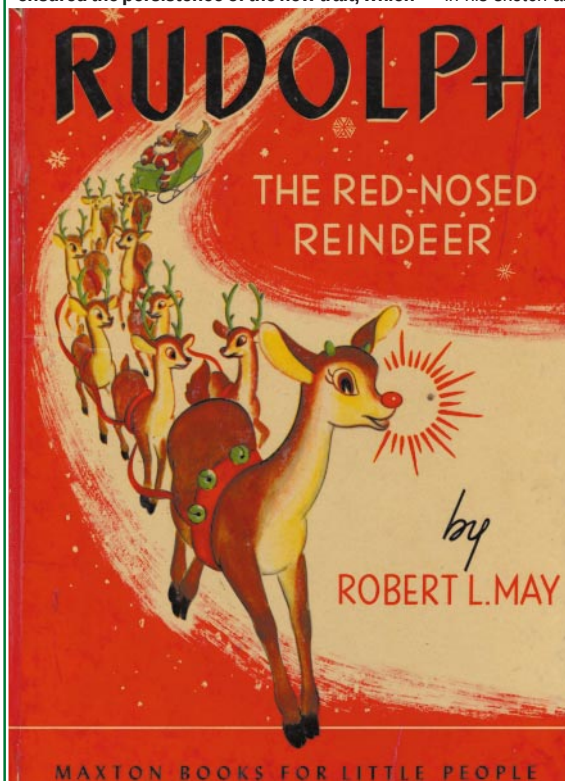


Figure 1 The first known recorded image of Rudolph, the red-nosed mutant of the reindeer, *Rangifer tarandus rubens*.

The living Earth

James Lovelock

Imagine a science-based civilization far distant in the Galaxy that had built an interferometer of such resolving power that it could analyse the chemical composition of our atmosphere. Simply from this analysis, they could confidently conclude that Earth, alone among the planets of the Solar System, had a carbon-based life and an industrial civilization. They would have seen methane and oxygen coexisting in the upper atmosphere, and their chemists would have known that these gases are continually consumed and replaced. The odds of this happening by chance inorganic chemistry are very long indeed. Such persistent deep atmospheric disequilibrium reveals the low entropy characteristic of life. They would conclude that ours was a live planet — and the presence of CFCs in the atmosphere would suggest an industry unwise enough to have allowed their escape.

As part of NASA's planetary exploration team in 1965, thoughts such as these led me to propose atmospheric analysis for detecting life on Mars. I also wondered what could be keeping Earth's chemically unstable atmosphere constant and so appropriate for life, and what kept the climate tolerable despite a 30% increase in solar luminosity since the Earth formed. Together, these thoughts led me to the hypothesis that living organisms regulate the atmosphere in their own interest, and the novelist William Golding suggested Gaia as its name. Although the concept of a live Earth is ancient, Newton was the first scientist to compare the Earth to an animal or a vegetable. Hutton, Huxley and Vernadsky expressed similar views but, lacking quantitative

evidence, these earlier ideas remained anecdotal. In 1925 Alfred Lotka conjectured that it would be easier to model the evolution of organisms and their material environment coupled as a single entity than either of them separately. Gaia had its origins in these earlier thoughts, from the evidence gathered by the biogeochemists Alfred Redfield and Evelyn Hutchinson and from the mind-wrenching top-down view provided by NASA.

Although welcomed by atmospheric scientists, Earth scientists were cautious. Biologists, especially Ford Doolittle and Richard Dawkins, argued strongly that global self-regulation could never have evolved, as the organism was the unit of selection, not the biosphere. In time I realized that they were right — but still I thought, something keeps the Earth habitable. In 1981 I composed a model of dark- and light-coloured plants that competed for growth on a planet in progressively increasing sunlight. My intention was not to make a blueprint for the Earth, but a model to show that Gaia is consistent with natural selection. This 'Daisyworld' regulated its temperature close to that fittest for plant growth and — unusually for an evolutionary model made from coupled differential equations — it was stable, insensitive to initial conditions and resistant to perturbation. Daisyworld is darwinian, but the evolution of the organisms and the evolution of temperature proceed as a single, coupled process. The model was much criticized, but so far has resisted falsification. It was easy to show that Daisyworld tolerates 'cheats' — daisies that grow but offer nothing towards self-regulation. Other critics claimed that daisies would adapt to changing temperature and therefore simply

Gaia

Organisms and their environment evolve as a single, self-regulating system.

track temperature change, not regulate it. But the restraining function connecting growth with temperature is not negotiable; chemistry, not biology, sets its constants.

At this stage, the Gaia theory was missing plausible control mechanisms. The first discovered was a biological process that redressed the imbalance of the nutritious elements sulphur and iodine — these are abundant in the oceans, but deficient on the land surface. It was widely assumed that hydrogen sulphide and sea salt aerosol drifted from the ocean to the land. In 1971 I discovered that methyl iodide and dimethyl sulphide were ubiquitous in the Atlantic surface waters, and from my measurements Peter Liss calculated their fluxes in 1974. He argued that these biogenic gases were the main carriers of the natural elemental cycles of sulphur and iodine.

Then in 1982, the geochemists James Walker, P. B. Hayes and Jim Kasting suggested that the weathering of calcium silicate rock could regulate carbon dioxide and climate. Greater warmth leads to more rainfall and a faster removal of carbon dioxide from the atmosphere by rock weathering, which provides a negative feedback on temperature. This plausible mechanism is by itself too small to account for the observed rate of weathering. Organisms on the rocks and in the soil bring it to life as a Gaian mechanism; their growth varies with temperature and their presence amplifies the rate of weathering.

In 1986, there was the awesome discovery by Robert Charlson, James Lovelock, Meinrat Andreae and Steven Warren of a connection between biogenic dimethyl sulphide gas — the product of ocean algae — its oxidation in the atmosphere to form cloud condensation nuclei, and the subsequent effect of the clouds formed on climate. We wondered whether this could be a Gaian regulatory mechanism through the feedback between climate change and algal growth.

By the end of the 1980s there was sufficient evidence, models and mechanisms, to justify a provisional Gaia theory. Briefly, it states that organisms and their material environment evolve as a single coupled system, from which emerges the sustained self-regulation of climate and chemistry at a habitable state for whatever is the current biota.

Like life, Gaia is an emergent phenomenon, comprehensible intuitively, but difficult or impossible to analyse by reduction — not surprisingly it is often misunderstood. A simple automatic mechanism, like a



Our planet in perspective: Gaia theory explains the constancy of our unstable atmosphere.

NASA/BETTMANN/CORBIS

thermostatically controlled oven, requires a sensor to measure the difference between the ambient temperature and the set point of regulation, and an amplifier to magnify this difference and apply it as negative feedback to oppose unwanted change. Living systems rarely work in this simple way; they require positive as well as negative feedback for homeostasis, and a restraining function replaces the simple manual set point. This function allows regulation within a physiologically acceptable range, instead of at a single set value. Andrew Watson and other critics have assumed that to be Gaian, a planet must regulate near perfectly—but physiological systems may perform no better than is needed. No one doubts that humans are in thermostasis, yet our core temperatures range from 35 to 40 °C and our extremities from 5 to 45 °C. This may appear imprecise, but it serves us well. For the past ten million years the Earth's average surface temperature has covered a similar range between 11 and 16 °C. This is not evidence of incompetent regulation—it is sufficient to sustain the Earth system. The occasional failure of the Earth to regulate efficiently—as in the present interglacial—resembles the physiological condition of a fever where positive feedback dominates.

Gaia theory does not contradict darwinism, rather it extends it to include evolutionary biology and evolutionary geology as a single science. In Gaia theory, organisms change their material environment as well as adapt to it. Selection favours the improvers, and the expansion of favourable traits extends local improvement and can make it global. Inevitably there will be extinctions and losers, winners may gain in the short term, but the only long-term beneficiary is life itself. Its persistence for over three billion years in spite of numerous catastrophes, internal or external, lends support to the theory. I have never intended the powerful metaphor 'the living Earth' more seriously than the metaphor of 'the selfish gene'. I have used it, along with my neologism geophysiology, to draw attention to the similarity between Gaian and physiological regulation.

I was pleased when Stephen Schneider persuaded the distinguished American Geophysical Union to devote their 1988 Chapman Conference to Gaia, but disappointed when too many of those who attended argued against the discarded Gaia hypothesis of the 1970s, seemingly unaware that the theory had been revised. I suspected that few would take Gaia seriously until eminent scientists approved it publicly. In 1995 I started dialogues with John Maynard Smith and William Hamilton. Both of them were prepared to discuss Gaia as a scientific topic, but neither of them saw how planetary self-regulation could evolve through natural selection. Even so, Maynard Smith gave unstinted support to my colleague Tim Lenton when he wrote a seminal article in *Nature*.

Table 1 Some predictions from Gaia

Prediction (year)	Test and result
Mars lifeless from atmospheric evidence (1968).	Viking Mission (1977). Strong confirmation.
That elements are transferred from the ocean to the land by biogenic gases (1971).	Dimethyl sulphide, dimethyl selenide and methyl iodide found (1973, 2000).
Climate regulation through biologically enhanced rock weathering (1981).	Microorganisms found greatly to increase the rate of rock weathering.
That Gaia is aged (1982).	Generally accepted.
Climate regulation through cloud albedo control linked to algal gas emissions (1987).	Still under test.
Archaean atmospheric chemistry dominated by methane (1988).	Still under test but tending to be accepted.
Oxygen has not varied from $21 \pm 5\%$ for the past 200 million years (1989).	Still under test.
Boreal forests regulate their regional climate in a Daisyworld manner (1988).	Now part of global climate modelling.
Biodiversity is a necessary part of planetary self-regulation (1992).	Tested by models, but not yet in the field.
That the current interglacial is an example of system failure in a physiological sense (1996).	Still controversial.

Hamilton wondered, in a joint paper with Lenton, if the need of organisms to disperse was the link that connected ocean algae with climate. In a 1999 television programme, Hamilton said: "Just as the observations of Copernicus needed a Newton to explain them, we need another Newton to explain how darwinian evolution leads to a habitable planet."

Then the ice began to melt. In 2001, at a conference in Amsterdam—at which four principal global change research programmes were represented—more than a thousand delegates signed a declaration that started with the statement: "The Earth System behaves as a single, self-regulating system comprised of physical, chemical, biological and human components."

Gaia theory is fruitful and makes successful or useful predictions (see Table 1). More than this, it enlightens our view of Earth system science and the environment. Importantly, as Lynn Margulis has insisted, it draws our attention to the microorganisms, which are the biological infrastructure of the Earth. Microorganisms filled the whole biosphere for the greater part of life's history and they are still vital for effective planetary regulation.

A major achievement of Gaia has been the change in style of Earth system models. Climatologists, notably Peter Cox, Richard Betts and John Schellnhuber and colleagues, now include a responsive biota in their models of future climates, and their contributions have added realism to the predictions of the 2001 Intergovernmental Panel on Climate Change third assessment report.

As the Earth ages, the Sun's heat ineluctably intensifies; in approximately one billion years the Earth will pass the limit of climatic stability and irreversibly return to

inorganic chemistry. Moreover, as it grows older the Earth system weakens, and before long a large planetesimal impact may throw our planet prematurely into its final hot, dry state. A few thermophiles in oasis ecosystems might survive, but we could never recapture the abundant life and lush environment we now enjoy. The Earth system is elderly and we should treat it with respect and care.

Gaia theory reconciles current thinking in evolutionary biology with that in evolutionary geology. It extends, not contradicts, Darwin's vision, just as relativity enhances, not denies, Newtonian physics. The theory is provisional, but provides an intellectual habitat where understanding of the Earth can evolve and grow. Perhaps its greatest value lies in its metaphor of a living Earth, which reminds us that we are part of it and that human rights are constrained by the needs of our planetary partners. ■

James Lovelock is at Green College, Woodstock Road, Oxford OX2 6HG, UK.

FURTHER READING

- Kump, L. R., Kasting, J. F. & Crane, R. G. *The Earth System* 2nd edn (Pearson Prentice Hall, Upper Saddle River, New Jersey, 2004).
- Hamilton, W. D. & Lenton, T.M. *Ethol. Ecol. Evol.* **10**, 1–16 (1998).
- Lenton, T. M. *Nature* **394**, 439–447 (1998).
- Lovelock, J. E. *The Ages of Gaia* (Oxford Univ. Press, Oxford, 2000).
- Margulis, L. *The Symbiotic Planet* (Phoenix, London, 1999).
- Schellnhuber, H.-J. & Wenzel, V. (eds) *Earth System Analysis* (Springer, New York, 1998).
- Turney, J. *Lovelock and Gaia* (Icon, Cambridge, 2003).
- The Amsterdam Declaration on Global Change* online at http://www.sciconf.igbp.kva.se/AMS_DECLARATION.pdf

From the grapevine

Summarize yourself in the form of a title of a paper in Nature.

Stochastic events govern individual lineage and fate in the MD–PhD physician–scientist career pathway.

Who has been the most important mentor in your career?

The key mentoring steps in my career came as a result of the combinatorial effect of unique individuals at the University of Texas Southwestern Medical Center in Dallas and the University of California, San Diego (UCSD). At Parkland Hospital in Dallas, clinical rounds were made each day with a star-studded cast of physician–scientists assembled by Don Seldin, who emphasized that it was not only possible to mix state-of-the-art clinical medicine with world-class science, but that the two were synergistic. This was at a time when the poles of science and clinical medicine were starting to diverge at many other medical schools. As a fledgling physician–scientist, I benefited from the unique collaborative and collegial environment in La Jolla, where the cross-institutional barriers are few, and the scientific–clinical bridges are many.

What literary character would you employ as a postdoc?

John Le Carré's spymaster, George Smiley.

What's your favourite conference destination, and why?

The Salk Institute in San Diego: architecturally breathtaking, scientifically exciting and within walking distance of the UCSD campus.

You have the audience in your hands, but some smart-alec asks you the killer question you have no idea how to answer. What's your response?

Invite them to an interview for a position in my lab.

What book is currently on your bedside table?
The Impressionist by Hari Kunzru.

What music heads the playlist in your car or lab?
Bob Dylan's Tangled Up in Blue.

The job of Captain on the Starship Enterprise in Star Trek has become vacant. Nominate any real person, living or dead, for the post.

Roger Revelle or Jonas Salk. Respectively, they built up UCSD and the Salk Institute from raw ideas to a leading public university and independent research institute within 20 years.

Where and when would you most like to have lived or worked?

Same time and same place, but oh, to be 20 years younger...

What was the worst/most memorable comment you ever received from a referee?

"This grant is a high-risk fishing expedition."

You are on a plane behind two students obviously going to the same conference, who start to talk about your work. What do you do?

Put on my Bose headset and hope for the best.

What do you most dislike about having research published?

Work that addresses complex aspects of human disease usually has multiple authors; there needs to be an improvement in the assignment of credit to each. This type of work requires a 'village' that extends beyond the first and senior authors.

What do you do to relax?

Wine works well with my Chinese alcohol dehydrogenase deficiency.

What would you have become, if not a scientist?
A wine master.

What previously under-recognized sport or pastime should be included in the Olympic Games?
Fly fishing — it requires skill and knowledge of where, when and what, rather like finding key genes within human disease pathways.

If you were reborn as a comic-book superhero, what would be your superhuman power?
Functioning normally without sleep.

Under what conditions do you have your greatest and most inspired ideas?
Drinking wine.

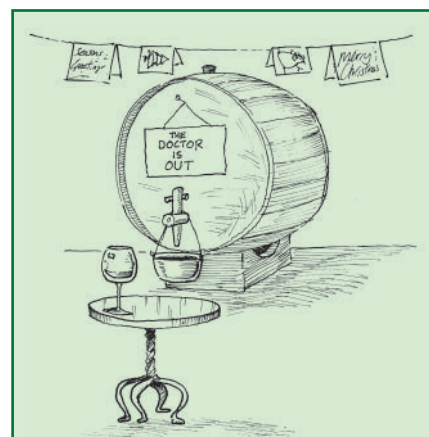
Is there a 'tyranny of reductionism' in how scientists are trained today?

I believe that innovation lies in the integration of ideas from many fields. We now can think of approaching human biology with an arsenal of tools that transcend reductionism. In a sense, we have been training Olympic athletes to compete in single events, when now the future may belong to the decathletes who excel in specific events, but are trained broadly enough to function well in others. To do this, a new interdisciplinary infrastructure for training in the biomedical sciences will be needed.

What's the one thing about science that you wish the public understood better?

That the future of medicine is science, science and more science.

You've just been told (in confidence) that the world will end tomorrow. What do you do next?
Put on my Bose headset and pour a glass of the best wine I can find.



Ken Chien

Ken Chien is the director of the Institute of Molecular Medicine at the University of California, San Diego, and a professor in the university's department of medicine and at the Salk Institute (adjunct). His hobbies include wine tasting, oriental art and losing to his daughters at tennis.

What overlooked or underrated discovery really changed the science in which you work?

The ability to miniaturize *in vivo* physiological assays for complex aspects of heart function in the living mouse, which normally has a heart rate in excess of 500 b.p.m. and an aortic diameter of less than 1 mm.

Which field in science deserves more funding?
Studying the origins of humanity using the tools of modern biology.

What would you change about Nature?
Be wary of starting too many more journal spin-offs.

What music would you have played at your funeral?

The second movement of Bach's double violin concerto.

I detect a certain oenological theme in your answers — and yet you are a cardiologist. Is drinking wine really good for you?

The development of a great wine is a prime example of conquering biological complexity. Although the raw ingredients are related to climate, geography, pedigree and cultivation, real achievement often requires a unique combination of art, science and the human skills of taste, discrimination and passion. Around the globe, where there is wonderful wine, there are usually wonderful people.

This happens to be the last Lifelines column. Do you have any further wisdom you'd like to impart, before we turn off the lights?
Without mentors, there would be no lifelines.

Thank you

A magnificent seven

Of the 325 News and Views articles published this year, seven are singled out for special attention. They illustrate the great job that scientists can do in communicating and commenting on new research.

One of the duties of an editor on a scientific publication is to take a succinct, clear and informative piece of writing and, through the leaden application of house rules, turn it into unreadable porridge. That's the view of the sceptical scientist and there's occasional truth in the observation. The better — and less wearing — editorial course for such an article is simply to publish it, tweaked little, if at all. Writers of such pieces are to be treasured by editors and readers alike.

Hence this celebration of seven authors who wrote outstanding News and Views articles for *Nature* in 2003: David Wark, Philip N. Benfey, S. Blair Hedges, Steve Blinkhorn, John Harte, Toren Finkel and Len A. Fisk. This is a competition that the contenders did not enter. Invidious though it is in that circumstance to choose between very fine and fine examples of exposition, the first three are winners and the other four 'highly commended'. The authors have respectively earned themselves magnums and bottles of champagne.

The seven have been chosen not by a panel of the great and the good in science, but by the three News and Views editors at *Nature*. The thinking behind that stems from a sceptic's view from the editing side of the fence. It is not clear from a published article how much of the structure and prose, good or bad, is due to the author and how much to the editor. True, cer-

tain scientists become justifiably well known for the clarity of their writing. But when the object to be gauged is communication of content, as much as content itself, a judgement of any given article cannot really be made except from its raw form.

Thus the choice of judges, who also came cheaply. And thus the choice of these seven articles, which were published much as written. From a short list of 30 or so, seven displayed the relevant qualities (a number that also makes for an eye-catching title).

News and Views articles, and similar writing in other journals, occupy unusual publishing ground. In both style and content, they lie between journalism and the scientific review literature. Sometimes the principles of the two conflict. But when liveliness and expert opinion come together well, they make for a very happy combination. Over the years there have been some fine and prolific exponents, such as Robert May (now president of the Royal Society) and Jared Diamond (a Pulitzer prizewinner).

The News and Views section first appeared in *Nature* in the issue of 2 January 1926, but curiously enough as 'Views and News'. There was evidently a quick rethink, because in the next issue the words were swapped into their more logical order. Only in the mid-1960s, however, did the section take its modern form, the hallmark being

scientists writing about the work of other scientists in newsy fashion. Of the star authors featured here, five were commenting on papers published in *Nature*, and two on work that appeared elsewhere. Their articles have a variety of virtues. One is to engage the reader from the beginning. A second is the effective use of metaphor or analogy. A third is a certain narrative thread, avoiding discursive detail. A fourth is explaining technicalities, where essential, and avoiding them where not. And a fifth is the appropriate visual support from graphics or photos.

All this is good as far as it goes, but each of these pieces offers more. They provide context and opinion on the work under discussion, so living up to the second part of the apparently cutesy name of News and Views. Context and opinion come in the form of ideas for further investigation, comparison with other bodies of evidence or theory, and comment on the limitations of the new work. Such critical appraisal is a central service offered by authors writing for a broad scientific public — or any public, come to that.

Congratulations and thanks to this particular magnificent seven. But thanks, too, to all of the other writers who have taken the time and trouble to produce News and Views articles in 2003. Once the hangover from the holiday season is over, we look forward to more of the same in 2004.

Tim Lincoln

Champagne writing

These articles can be seen at www.nature.com/nature/newsandviews/magnificent7

● **David Wark** "Particle physics: Now you see them, now you don't" (421, 485–486)

Readable and thoughtful account of the latest evidence that types of neutrino can interchange. Holds the attention despite its length.

● **Philip N. Benfey** "Molecular biology: MicroRNA is here to stay" (425, 244–245)

Goes beyond the main paper under discussion in surveying previous work, neatly stepping across the bog of abbreviations that makes writing on cell and molecular biology so tough.

● **S. Blair Hedges** "Biogeography: The coelacanth of frogs" (425, 669–670)

Does full justice to a cracking story, which at first sight seems of specialist interest only, taking in three disparate themes.

● **Steve Blinkhorn** "Neuroscience: Of mice and mentality" (424, 1004–1005)

Fizzy but properly cautious appraisal of the possible existence of a mouse version of IQ, all done in under a page of text.

● **John Harte** "Ecology: Tail of death and resurrection" (424, 1006–1007)

Brisk assessment, spiced with philosophical asides, of the standing of the neutral theory in ecology: not an easy topic.



● **Toren Finkel** "Ageing: A toast to long life" (425, 132–133)

History used with wit to top and tail a succinct account of one angle on lifespan extension in yeast: never gets lost in the details (and includes a joke).

● **Len A. Fisk** "Planetary science: Over the edge?" (426, 21–22)

Balanced, well-paced account of the adventures of Voyagers 1 and 2 at the edge of the Solar System, and the associated controversy.

BETTMANN/CORBIS

Archaeology

Art of the ancients

Anthony Sinclair

One might imagine that the first examples of art would be simple and crude. New finds bolster the evidence that modern humans were astonishingly quick in developing their artistic skills.

Excavations at the cave site of Hohle Fels in southwestern Germany have unearthed three small animal figurines, as Nicholas Conard reports on page 830 of this issue¹. The figurines, none longer than 2 cm, were carved out of mammoth ivory sometime between 30,000 and 33,000 years ago by some of the first modern humans to colonize Europe. One is shaped like a bird, another like the head of a horse, and the third seems to be half-man, half-animal.

This find is an important addition to a group of more than 20 ivory figurines that have been found at four sites in the Ach and Lone Valleys: Vogelherd, Geissenklösterle, Hohlenstein-Stadel and Hohle Fels. Without question, they are the oldest body of figurative art in the world — pieces that show a coherent set of manufacturing techniques and themes for representation. Alongside the figurines found at each of these four sites are the remains of waste from ivory, bone and stone working. At these four sites, could we be looking at the oldest artists' workshops?

The study of early art has been plagued by our desire to see this essentially human skill in a progressive evolutionary context: simple artistic expressions should lead to later, more sophisticated creations. We imagine that the

first artists worked with a small range of materials and techniques, and produced a limited range of representations of the world around them. As new materials and new techniques were developed, we should see this pattern of evolution in the archaeological record. Yet for many outlets of artistic expression — cave paintings, textiles, ceramics and musical instruments — the evidence increasingly refuses to fit. Instead of a gradual evolution of skills, the first modern humans in Europe were in fact astonishingly precocious artists.

For example, before we were able to date directly the classic cave paintings of southwestern France and Spain, eminent archaeologists had devised evolutionary schemes that ordered the works, from the first charcoal animal drawings to the more recent multicolour animals drawn with a clear sense of perspective at famous sites such as Lascaux and Altamira. And yet the beautiful multicolour horses, lions and mammoths at the Grotte Chauvet in France, discovered in the early 1990s and dating from 32,400 years before present, are now thought to be the oldest examples of cave art in the world (Fig. 1).

The oldest evidence for the use of textiles and clay — at Pavlov and Dolni Vestonice in

the Czech Republic and some 26,000 years old — does not suggest crude techniques or a poor knowledge of materials. Among the textiles, there are a number of distinct patterns, and also some finely made clay figurines. Furthermore, the evidence of several thousands of small pieces of ceramic has suggested to some that humans could deliberately manipulate their raw materials to explode ceramics in their kilns on command — perhaps the earliest form of pyrotechnics.

At Geissenklösterle, Germany, a fragment of a bone pipe, of the musical sort, has been found. This pipe was made from the radius bone of a swan and has three clear finger holes. There are also more than twenty specimens of musical pipe of the same sort and age from Isturitz in France. Microscopic examination suggests that they may have been reed-voiced instruments, like a modern oboe, and that the finger holes have been chamfered to increase the pneumatic efficiency of the finger seal: simple whistles they are not. Such evidence of complexity is used to argue that these cannot be the first musical pipes, even though they are the oldest in the archaeological record².

All of these finds could still be accommodated in our old evolutionary schemes, if we imagine that the first steps in the discovery and use of these skills were undertaken using more perishable materials; or perhaps earlier examples were simply not discarded in the sites that have been excavated so far. There are examples of unquestionable art before modern humans — such as the engraved stone cortex from the site of Quneitra in Israel, dated to 50,000 years before present — but these can be counted on the fingers of one hand. The argument in favour of fast-developing artistic skills in modern humans is strong, and certainly one that I find convincing.

Returning to Conard's finds at Hohle Fels¹ and those at neighbouring sites in Germany, there is more to challenge our perceptions of the past. These figurines are not 'art' as we know it today — they are personal possessions, and the evidence of very discrete activities. Each figurine shows a series of incised lines, or sometimes crosses, along its back or sides; polishing from constant handling is clear to see³. They are found in contexts that suggest they might have been cached for later use, or deliberately buried.

The animal species usually represented — mammoth, bear and lion — are not ones that would have been eaten (a common feature with this early art). The postures of these figurines, and their ears and eyes, reveal a close attention paid to aggressive animal behaviour. They may be material expressions of the shared personal qualities of humans and animals, as is perhaps indicated by the small half-human, half-animal figurine

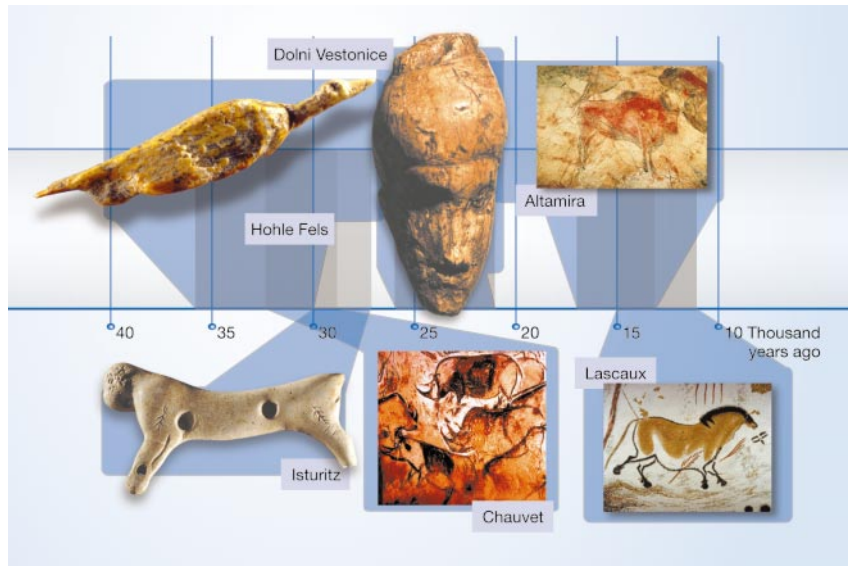


Figure 1 Prehistoric art. Three figurines from Hohle Fels, Germany¹ — including the bird-like figure shown here — are the oldest examples of figurative art that have been found in Europe. Examples of carving at Isturitz, France, and at Dolni Vestonice in the Czech Republic, and of cave painting at Grotte Chauvet and Lascaux, France, and at Altamira in Spain, show remarkable sophistication. Modern humans seem to have developed artistic skill quickly.

from Hohle Fels¹, and the larger half-lion, half-human figurine found earlier at Hohlenstein-Stadel. Or they may be expressions of the shared social qualities of single- and group-living species.

The Victorian idea of progressive evolution has been a very persuasive metaphor for explaining change in the archaeological record, particularly over a time of biological change in the human species. Yet the archaeological evidence is now forcing us to come

up with new timescales for cultural change and innovation. This is a challenge that makes the smallest finds of archaeology as important as the largest. ■

Anthony Sinclair is in the Department of Archaeology, University of Liverpool, Liverpool L69 3BX, UK.

e-mail: a.g.m.sinclair@liv.ac.uk

1. Conard, N. J. *Nature* **426**, 830–832 (2003).
2. d'Errico, F. et al. *J. World Prehist.* **17**, 1–70 (2003).
3. White, R. *Annu. Rev. Anthropol.* **21**, 537–564 (1992).

Cell biology

Earthworms and lipid couriers

Sean Munro

Lipids can hop between cellular compartments without using the transport vesicles that carry proteins. A key molecule involved in conveying the lipid ceramide has at last been uncovered.

Slice open any eukaryotic cell and you will find it packed with membrane-bounded compartments, each with a characteristic set of resident proteins and lipids. Many of these molecules begin their lives in the endoplasmic reticulum (ER), a large network of intracellular membranes. The process by which newly made proteins are moved from the ER to their final destinations in the cell is well known: they are transported in small containers, or 'vesicles', of membrane that pinch off from the ER and then move to, and fuse with, their target membrane.

Some lipids, however, can bypass this transport process, instead moving directly from the ER to other compartments. The mechanisms involved in this non-vesicular transport are poorly understood, but on page 803 of this issue Hanada and co-workers¹ describe the identification of CERT — a protein that seems to convey the lipid precursor ceramide from the ER to the Golgi apparatus. The parts of CERT that allow it to act as a courier are seen in several families of proteins of unknown function, so the new work also highlights candidates for involvement in other lipid-transport events.

Ceramide is the backbone molecule of the sphingolipids — abundant components of the outer membranes of eukaryotic cells. It is synthesized in the ER, but the addition of headgroups to complete the synthesis of sphingolipids takes place in the Golgi apparatus. Conversion of ceramide into sphingolipids is known to continue when vesicle transport from the ER to the Golgi is blocked^{2,3}, so ceramide clearly does not rely on vesicles to reach the Golgi. But quite how it is transported has been unclear.

Hanada and colleagues' route to answering this question¹ started with some help from an unusual source: the humble earthworm.

This creature's bodily fluid contains a toxic protein known as lysenin, which binds to sphingomyelin — the most abundant of the sphingolipids⁴. Previously, lysenin was used to isolate mutant mammalian cells that are resistant to the toxin because they have less sphingomyelin than usual⁵. Some of these mutant cells have defects in making ceramide. But in one mutant, called LY-A, ceramide is made but its delivery to the Golgi is greatly reduced.

Hanada and co-workers set about identifying the gene that is mutated in these LY-A cells. To do so, they introduced different wild-type genes into the cells until they found one that could restore wild-type levels of sphingomyelin synthesis. They named this gene CERT. To confirm that they had fished out the correct gene, the authors examined CERT from the LY-A mutant cells, and found that, as expected, it carried a mutation — with the result that the amino acid glycine at position 67 of the encoded protein was replaced by glutamate (a replacement symbolized by G67E).

So what role might CERT have in ceramide transport? This proved to be one of those happy occasions when the amino-acid sequence of a newly discovered protein gave a strong clue to how it might act (Fig. 1a). CERT is predicted to be a cytoplasmic protein that contains a START domain — a structural region that in several other proteins forms a deep lipid-binding pocket⁶. In addition, CERT has two motifs known to be involved in subcellular targeting. The first is a pleckstrin homology (PH) domain, which is a member of a family of closely related PH domains that bind to the Golgi apparatus⁷. The second domain is a FFAT motif (comprising two phenylalanine amino acids, 'FF', in an acid tract). This motif is found in several cytoplasmic proteins involved in lipid metabolism, and it binds to VAP,

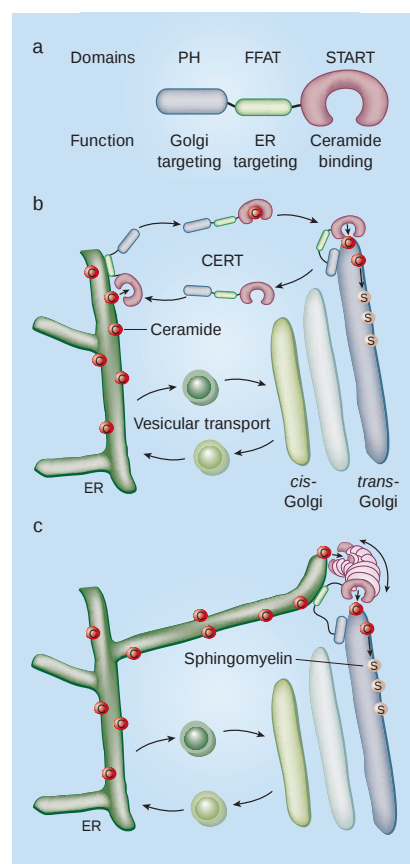


Figure 1 Conveying ceramide from the endoplasmic reticulum (ER) to the Golgi. a, Hanada et al.¹ have identified the CERT protein, which contains a PH domain that binds to the Golgi apparatus⁷, a FFAT motif that in other proteins binds to the ER protein VAP^{8,11}, and a START domain that binds ceramide. b, CERT might work by extracting ceramide (C) from the ER and then diffusing with it through the cytosol to deliver it to the Golgi. There, ceramide is converted to sphingomyelin (S) by the enzyme sphingomyelinase, located on the far (trans) side of the Golgi¹². c, Alternatively, CERT might bind ER and Golgi membranes simultaneously, allowing the START domain to mediate rapid lipid transfer. Contact sites between the ER and trans-Golgi have been observed by electron microscopy¹³. Proteins, meanwhile, are transported between the ER and the near (cis) side of the Golgi in vesicles.

a protein anchored in the ER membrane⁸.

This combination of a lipid-binding domain and targeting domains for both ER and Golgi suggested that CERT acts directly as a ceramide carrier that shuttles between the two organelles. Hanada and colleagues found that CERT's properties are entirely consistent with this model. First, CERT can extract ceramide, but not other lipids, from artificial membranes, and this activity requires the START domain. Second, CERT can catalyse the transfer of ceramide

between artificial membranes, indicating that it can release ceramide at an acceptor membrane. Third, the G67E mutation (which renders CERT defective in the LY-A mutant cells) is found in the PH domain, and this mutation prevents CERT from binding to the Golgi — suggesting that targeting to the Golgi is important for CERT function. Finally, wild-type CERT can be added back to LY-A extracts to restore ceramide transport *in vitro*, and this activity is lost if any one of the START, PH or FFAT domains is deleted.

So, a clear model emerges in which CERT carries ceramide across the cytosol by binding to VAP on the ER, extracting a ceramide molecule, and then binding to the Golgi (Fig. 1b). As always, however, several questions remain. For instance, the cell must ensure that ceramide is moved from the ER to the Golgi but not back again, and it is not clear whether this directionality is provided simply by the ceramide's being made in the ER and then, on arrival in the Golgi, being converted to sphingomyelin (which is not a substrate for CERT). Another issue is why the *in vitro* transport assay requires ATP⁵. This molecule, as well as being the main cellular energy store, also supplies phosphate groups for protein and lipid modification. The Golgi-specific PH domains bind to the membrane lipid phosphatidylinositol-4-phosphate⁷, so perhaps ATP is required for producing this lipid. But it may also be that the cycle of ceramide binding and release is regulated by phosphorylation.

These detailed issues aside, the discovery of CERT also has broader implications for the cell biology of lipids. As noted above, there are several well-conserved eukaryotic proteins that have FFAT or Golgi-specific PH domains, and in some cases putative lipid-binding domains^{8,9}. These proteins' functions are unclear, but they are clearly excellent candidates for transporting other lipids between organelles. Moreover, ceramide has been suggested to have an additional role in transmitting signals into the cell from the plasma membrane¹⁰. So ceramide transport to intracellular organelles might be important in attenuating these signal-transduction processes. Finally, the presence in CERT and in other proteins of targeting domains for more than one organelle raises the possibility that these proteins might not simply be shuttles. Instead, they could actually form physical bridges between organelles, with rapid transport mediated by oscillation of the lipid-binding domains (Fig. 1c)⁸.

There is still much to be learnt about the non-vesicular movement of lipids between membranes, but Hanada and colleagues' elegant combination of cell biology and lipid biochemistry has provided much food for thought.

Sean Munro is at the MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK.

e-mail: sean@mrc-lmb.cam.ac.uk

1. Hanada, K. *et al.* *Nature* **426**, 803–809 (2003).
2. Collins, R. N. & Warren, G. J. *Biol. Chem.* **267**, 24906–24911 (1992).
3. Kok, J. W., Babia, T., Klappe, K., Egea, G. & Hoekstra, D. *Biochem. J.* **333**, 779–786 (1998).
4. Yamaji, A. *et al.* *J. Biol. Chem.* **273**, 5300–5306 (1998).
5. Fukasawa, M., Nishijima, M. & Hanada, K. *J. Cell Biol.* **144**, 673–685 (1999).
6. Ponting, C. P. & Aravind, L. *Trends Biochem. Sci.* **24**, 130–132 (1999).
7. Levine, T. P. & Munro, S. *Curr. Biol.* **12**, 695–704 (2002).
8. Loewen, C. J., Roy, A. & Levine, T. P. *EMBO J.* **22**, 2025–2035 (2003).
9. Lehto, M. & Olkkonen, V. M. *Biochim. Biophys. Acta* **1631**, 1–11 (2003).
10. Adam, D., Heinrich, M., Kabelitz, D. & Schutze, S. *Trends Immunol.* **23**, 1–4 (2002).
11. Wyles, J. P., McMaster, C. R. & Ridgway, N. D. *J. Biol. Chem.* **277**, 29908–29918 (2002).
12. Pagano, R. E., Martin, O. C., Kang, H. C. & Haugland, R. P. *J. Cell Biol.* **113**, 1267–1279 (1991).
13. Ladinsky, M. S., Mastronarde, D. N., McIntosh, J. R., Howell, K. E. & Staehelin, L. A. *J. Cell Biol.* **144**, 1135–1149 (1999).

Organic chemistry

Aromatics do the twist

David M. Lemal

For nearly 40 years, organic chemists have been fascinated by the idea of aromatic molecules that have the topology of a Möbius strip. No such molecule has been isolated — until now.

The first example of a Möbius aromatic molecule to be synthesized and isolated is described by Ajami *et al.*¹ on page 819 of this issue. Although topologies resembling a Möbius strip (Fig. 1) are not unknown in chemistry, a stable aromatic molecule with this configuration had escaped fabrication because the obligatory twist could always be circumvented by a change in shape of the molecule. The authors have overcome this difficulty by assembling their molecule from two fragments, one of which was rigidly constructed in such a way as to enforce a twist when it was linked to the other.

To an organic chemist, an aromatic molecule is one that resembles the famous benzene molecule. This odd meaning for the description 'aromatic' came about early in the history of organic chemistry when it was discovered that certain derivatives of benzene have pleasant odours — wintergreen, anise, almond and cinnamon are examples. The benzene molecule consists of a planar, cyclic array of six CH fragments, each carbon atom contributing a *p*-type atomic orbital bearing one electron to a so-called π -system (Fig. 2a). This loop of overlapping *p* orbitals, in which the six electrons move freely, contributes much to the bonding in the ring, making the benzene ring a robust structural element that is present in myriad drugs, dyes, foods and plastics, as well as in all living systems.

In the 1930s Erich Hückel realized^{2,3}, with the help of quantum mechanics, that cyclic systems containing $4n+2$ π -electrons, where n is an integer, should be relatively stable — so not just 6-electron systems, but those with 2, 6, 10 or 14 electrons, and so on. His insight stimulated intensive research, beginning in the 1950s, that resulted in the creation of many such ring systems and the confirmation of their resemblance to benzene. In contrast, cyclic π -systems with $4n$ electrons (4, 8,



Figure 1 *Möbius Strip II* by M. C. Escher. The red ants marching along the surface of the twisted loop show that the Möbius strip has a single continuous surface — a topology that has now been realized in an aromatic compound by Ajami *et al.*¹ (M. C. Escher's "Moebius Strip II" © 2003 Cordon Art B. V. – Barn – Holland. All rights reserved.)

12, 16 and so on) were found to be either highly unstable or, at best, lacking the special stability of $4n+2$ compounds. Stability is an important criterion for aromaticity, but a variety of other measures, such as equivalence of bond length and magnetic properties, are also used to detect and quantify it⁴.

In 1964, not long after the first large annulenes (hydrocarbons with cyclic π -systems) had been synthesized, Edgar Heilbronner began to wonder about their π -orbital topology. He suggested that rings of about 20 members or more could incorporate a 180° twist without being strained: the π -system would assume the topology of a Möbius band⁵ (Fig. 2b). The cyclic π -systems known at that time could be represented schematically by a circular band with two sides and two edges, but the twisted Möbius band would have just a single surface and a single edge.

Heilbronner used simple quantum mechanical methods to predict the properties of Möbius π -systems. He concluded that, although the half-twist strongly destabilized $4n+2$ π -systems, $4n$ systems were actually stabilized, if the weakening effect of diminished π -orbital overlap caused by the twist were ignored. In his calculations, these two effects in $4n$ systems exactly cancelled, so the topological switch from simple loop to Möbius strip would leave the system unchanged in energy. Thus, there seemed to be no driving force that would encourage annulenes to twist, and many chemists may have regarded Heilbronner's imaginative insight as just an interesting academic curiosity.

Two years later, however, an important application of Heilbronner's idea was discovered. There is a large and important family of organic transformations called pericyclic reactions that proceed via a cyclic array of overlapping orbitals. For many of these reactions, two different orbital topologies for their (transient) transition states are geometrically feasible, leading to different products. The two topologies, it was realized^{6,7}, are Hückel (like benzene) and Möbius, and the observed product can be correctly predicted from the number of electrons in the cycle, $4n$ or $4n+2$. That number determines which topology enjoys aromatic character and thus which pathway requires less energy to traverse. Here the topologically different orbital arrays suffer a comparable amount of twisting, so the Möbius array is not energetically disadvantaged. Although nearly all textbooks of organic chemistry analyse pericyclic reactions in other terms⁸, the concept of Hückel and Möbius aromaticity offers a particularly simple, clear way of understanding them that is broadly applicable.

In recent years theoretical calculations more sophisticated than were possible when Heilbronner published his findings have

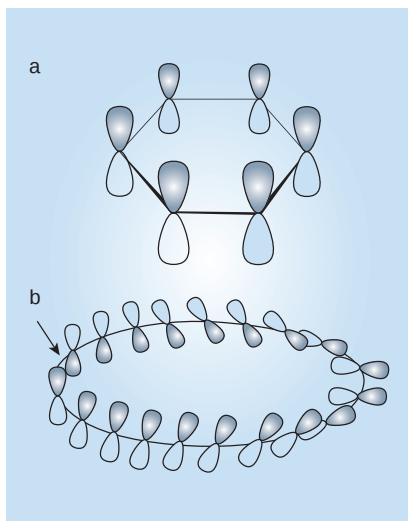


Figure 2 Twist in the tale. **a**, The π -system of benzene. The lobes represent the p orbitals contributed by each carbon atom. **b**, A hypothetical Möbius π -system. The arrow indicates the node that results from the 180° twist.

revealed that a variety of hydrocarbons with cyclic π -systems can assume Möbius conformations with aromatic character. As Ajami *et al.*¹ note, however, in every case but one the Möbius species are calculated to be less stable than others; the exception is a short-lived hydrocarbon ion that has not been directly observed. The challenge of creating a stable, isolable Möbius aromatic molecule has not

been easy. But Ajami *et al.* have accomplished that feat. A key feature of their design is a rigid structure incorporating a linear array of overlapping p orbitals, the end orbitals of which are forced to lie in a plane. To connect these termini to a second flexible p -orbital array requires the introduction of a twist. The twist takes three different forms, giving rise to three different Möbius structures, one of which has both Möbius topology and aromatic character.

Cleverly designed as it is, Ajami and colleagues' molecule does have some structural features that limit its aromaticity (orbital pyramidalization, benzo substitution and larger-than-optimal twist angles). But the fact that the authors have quite convincingly detected aromaticity in this molecule suggests that, in the future, Möbius molecules will be synthesized whose aromatic character far exceeds Heilbronner's modest expectations.

David M. Lemal is in the Department of Chemistry, Dartmouth College, Hanover, New Hampshire 03755, USA.

e-mail: david.m.lemal@dartmouth.edu

1. Ajami, D., Oeckler, O., Simon, A. & Herges, R. *Nature* **426**, 819–821 (2003).
2. Hückel, E. *Z. Physik* **70**, 204–286 (1931); **72**, 310–337 (1931); **76**, 628–648 (1932).
3. Yates, K. *Hückel Molecular Orbital Theory* (Academic, New York, 1978).
4. Garratt, P. J. *Aromaticity* (Wiley, New York, 1986).
5. Heilbronner, E. *Tetrahedron Lett.* **29**, 1923–1928 (1964).
6. Zimmerman, H. E. *J. Am. Chem. Soc.* **88**, 1564–1567 (1966).
7. Zimmerman, H. E. *Acc. Chem. Res.* **4**, 272–280 (1971).
8. Woodward, R. B. & Hoffmann, R. *Angew. Chem. Int. Edn* **8**, 781–853 (1969).

Biomechanics

Early birds surmount steep slopes

John R. Hutchinson

Even before they can fly, some young birds can run up vertical surfaces by using their wingbeats to add traction to their legs. Such behaviour may be relevant to understanding the origin of avian flight.

Birds, of course, typically use their wings to lift, then propel themselves through the air, and their legs to move on the ground. So the recent news that birds can use both sets of appendages together to scale inclined surfaces came as a surprise. The story began in 2001, with Terry Dial, 15-year-old son of Ken Dial of the University of Montana–Missoula Flight Laboratory. Dial junior noticed an unusual behaviour among the Flight Lab birds: he saw that some partridges could run up near-vertical surfaces, and that they seemed to use their flapping forelimbs (wings) and loping legs to do so.

The upshot was a study by Ken Dial¹, published in *Science* earlier this year, presenting evidence that chukar partridges (and three other close relatives of chickens) use their

wings much like the spoiler on an Indy 500 racer. The effect is to increase the load on their hindlimbs. In this way they can improve traction to get themselves up steep ascents. Dial dubbed this behaviour wing-assisted incline running (WAIR). He showed that the maximum slope the birds could traverse increased with age, wing feathering and surface friction. He also found that birds with full wings were able to use WAIR beyond the vertical, up to a slope of 105° , after which only free flight was viable. So without obvious specializations for climbing, ground birds could run straight up or even upside down. Not only was this a startling finding for those studying avian behaviour, natural history and flight biomechanics, but — as Dial argued¹ — it could tell us something about the evolution of flight.

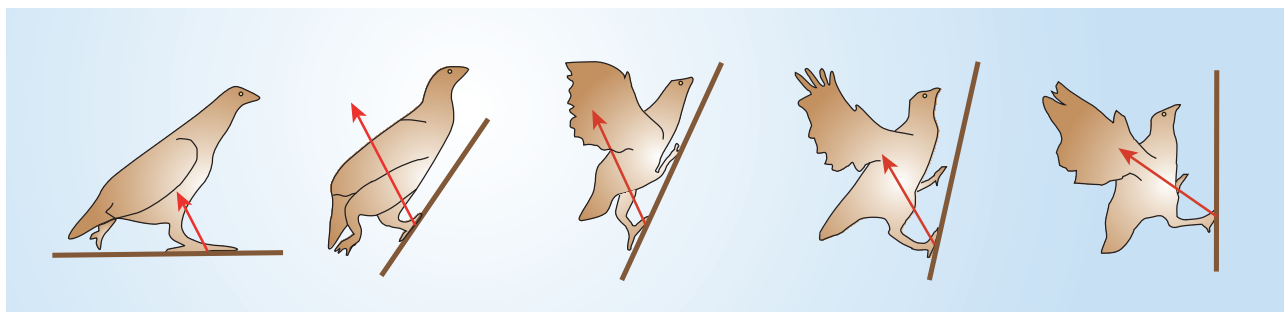


Figure 1 The mechanics of wing-assisted incline running, as described by Bundle and Dial². From left to right, chukars are shown running at inclinations of 0°, 60°, 70°, 80° and 90°. The arrows represent the direction and magnitude of the maximum ground reaction force incurred by the hindlimbs pushing against the substrate. The force remains directed away from the substrate (the birds are pushing down

to support the body), but on slopes the force increases and tilts forward (the birds are accelerating vertically along the incline). Even at the 90° inclination, added traction from the wingbeats allows the hindlimbs to exert large forces against the substrate, supporting the body and propelling it up the slope. (Figure modified from Fig. 5 of ref. 2, courtesy of Company of Biologists Ltd.)

A paper by Matt Bundle and Dial², which appears in *Journal of Experimental Biology*, now takes the fledgling study of WAIR to new heights. The authors asked two questions. How do the mechanics of the wingbeat compare between free flight and WAIR; that is, how flexible is wing function? And how much of a role do the hindlimbs have in WAIR — do they only provide support and stability, or propulsion as well?

To answer the first question, the movements and accelerations of birds flying freely at a 52° incline were compared with WAIR on a treadmill at the same incline. The treadmill pulled opposite to the direction of running, slowing the birds down so that a longer duration of WAIR could be filmed. Bundle and Dial deduced that the wingbeats accelerated the body towards the substrate in WAIR, perpendicular to the direction of travel, like a spoiler. By contrast, in free flight the wingbeats acted to accelerate the body more parallel to the direction of travel, as expected for a typical flight stroke. Hence wing function was quite flexible, able to shift from generating thrust and lift forwards and upwards during free flight to generating 'negative lift' towards the substrate in WAIR (Fig. 1).

The authors then used a ramp, designed to measure the forces exerted against the substrate, to determine the relative roles of the wings and legs in moving the body up the ramp at various inclinations. The chukars could climb 60° slopes without using their wings, then progressively used their wings more as the inclination increased. The percentage of vertical work (energy resisting the pull of gravity) produced by the hindlimbs declined suddenly between 60° and 70° inclination, from about 100% to 65% or less. The percentage of work that the forelimbs contributed increased with the slope, reversing the proportion of work done by the fore/hindlimbs to about 65%/35% at a slope of 90°. Surprisingly, even at a slope of 90° the hindlimbs continued to push against the substrate with

forces of more than 2.5 times body weight (Fig. 1). So the negative lift provided by the forelimbs enabled the hindlimbs to continue generating propulsion, presenting the novel case of two locomotor modules³ working in tandem.

This research is exciting purely because of the novelty of the WAIR behaviour. But the authors^{1,2} didn't stop there, and nor will I, for this work may illuminate the origin of flight in birds. Birds evolved from flightless theropod dinosaurs that were already capable runners. However, the evolutionary steps in flight function across this transition remain debatable. Dial recognized that even young birds with minute wings still benefited from these proto-wings during WAIR. Moreover, larger wings seemed to permit the ascent of steeper angles¹. This is a compelling solution for the evolutionary conundrum, 'What use is half a wing?'. The fossil record shows an evolutionary pattern of increasing plumage size and complexity from early flightless theropods to early birds^{4,5}, coincident with specializations of the muscles and skeleton that assembled the flight apparatus in a stepwise pattern^{3,6,7}. So WAIR might have been useful and usable at many steps along this series of transformations, enabling non-volant theropods to scurry up slanted surfaces for escape, predation or other reasons. When Dial first presented this work at the Society of Vertebrate Paleontology meeting in 2001, there was a commotion unlike any since the 'feathered' dinosaur *Sinosauropteryx*⁸ was reported there in 1996. Deservedly so — this is a stimulating development for a wide spectrum of fields.

There are plenty of issues yet to be explored, of course, which is a good thing for many researchers, including Bundle and Dial^{1,2}, who admit as much. The relative roles of inertial and aerodynamic forces remain unknown, as do the energetics of WAIR. I wonder how much hindlimb function changes between level running and WAIR, or how the third locomotor module, the tail, might influence WAIR. If WAIR is so

important in the natural history of ground (or other) birds, and if it is vital for birds to generate frictional forces to improve traction, might their feet be specialized accordingly? Likewise, it is not known how broadly distributed and crucial WAIR is for the thousands of species of extant birds, especially those such as tinamous, kiwis and ostriches that retain many ancestral avian traits (for example, spending a relatively short time in the nest after hatching, or frequenting more terrestrial than arboreal environments), although Bundle and Dial^{1,2} provide some tantalizing speculations. Finally, how closely are specific anatomical features of birds linked to functions integral to WAIR?

On the wish-list for the future would be the establishment of secure links between form and function for the group — Aves or Neornithes — that includes all extant birds and all descendants of their most recent common ancestor. If that could be done, then the more difficult historical questions of how WAIR evolved would become tractable because relationships between form and function could be traced across evolutionary lineages. Without such broader knowledge, it is uncertain how essential WAIR was for any extinct members of the theropod lineage, including the earliest birds. Regardless, this work will continue to stimulate research on flight and its evolution. The debate over whether flight originated in tree- or ground-living creatures is centuries old. The WAIR hypothesis has provided a biologically plausible alternative to that rather stale dichotomy. ■

John R. Hutchinson is at the Structure and Motion Laboratory, Royal Veterinary College, Hatfield, Hertfordshire AL9 7TA, UK.
e-mail: jrhutch@rvc.ac.uk

1. Dial, K. P. *Science* **299**, 402–404 (2003).
2. Bundle, M. W. & Dial, K. P. *J. Exp. Biol.* **206**, 4553–4564 (2003).
3. Gatesy, S. M. & Dial, K. P. *Evolution* **50**, 331–340 (1996).
4. Ji, Q. et al. *Nature* **410**, 1084–1088 (2001).
5. Prum, R. O. & Brush, A. H. Q. *Rev. Biol.* **77**, 261–296 (2002).
6. Jenkins, F. A. *Am. J. Sci.* **293**, 253–267 (1993).
7. Padian, K. & Chiappe, L. M. *Biol. Rev.* **73**, 1–42 (1998).
8. Chen, P.-J. et al. *Nature* **391**, 147–152 (1998).

Planetary science

Icy martian mysteries

Victor R. Baker

Both Mars and Earth have experienced ice ages in geologically recent times. Coincidence of the phenomenon on two planets will further the scientific quest to answer the question of how ice ages originate.

Among the grandest of mysteries about planet Earth is the origin of its ice ages and related climate change. Human civilization developed during a warm, 'greenhouse' climatic interlude of several thousand years within the overall 'ice-house' conditions of a major ice age that became most intense during the past two million years. On page 797 of this issue, Head *et al.*¹ present evidence that this same period coincides with an occurrence of martian ice ages. They develop a convincing case that these geologically recent ice ages were induced by episodes of higher tilt (obliquity) in the axis of Mars' planetary spin. The resulting warming of the martian polar caps increased the sublimation of water ice, the water vapour migrating to the then cooler lower latitudes where it was emplaced as ice-rich deposits.

Head *et al.* advance scientific arguments that are mainly geological in nature, not theoretical. By this I mean that the various indicators of past processes, consistent with appropriate physical principles, are used to infer the causes. This contrasts with scientific arguments that postulate principles, deduce predictions and then compare those predictions with observed features. Whereas theoretical arguments are corroborated by successful predictions, the geological approach derives its support from a full suite of evidence—in this case from the appropriate interpretation of landforms (Fig. 1) that indicate the presence of near-surface ice, emplaced or mobilized at the appropriate times over parts of the planet.

This means that the geologist is more an investigator than a theorist²: like a detective at a crime scene, the geologist relies on the evidence and knowledge of the operative processes to conclude what causes led to that evidence. The overall assemblage of evidence, and the explanatory surprises that it may generate ('consilience'), are used to suggest fruitful lines of inquiry. These tentative hypotheses are then subject to additional testing against new evidence. In other words, the geologist lets the planetary landforms tell their own 'story', just as the evidence at a crime scene reveals its story to an experienced detective.

Since the first orbital Mars missions of the 1970s, the geological evidence for ice- and water-related environmental change on Mars accumulated into a strong circumstantial case³. Nevertheless, it was the detection

of abundant near-surface hydrogen, made by neutron and γ -ray detectors aboard an orbiting spacecraft and published last year⁴, that provided the 'smoking gun': this indication of water ice at high martian latitudes finally convinced a majority of the planetary science community that Mars was indeed a water-rich planet. A physically plausible means for emplacing this water, consistent with what is known of the martian climate, is the exchange of water vapour between the atmosphere and the pore space of the martian soil⁵.

However, Head *et al.*¹ describe numerous attributes of landforms that indicate the direct emplacement of a mantle of ice and dust, possibly as dirty snow from the atmosphere. Such attributes include the regional smoothing of terrain by internally layered, and locally eroded, deposits a few metres thick; extensive, small-scale polygonal or patterned terrains (similar to the ice-wedge phenomena of Earth's polar regions); gullies and the mobilization of rocky debris on slopes (presumably by ice deformation); and even the likely remnants of glaciers, now mantled by dust or other debris (Fig. 1). Reinterpretation⁶ of the neutron and γ -ray measurements indicates that although sublimation and re-condensation of ice in pore spaces is probably active on Mars today, the measurements also require the emplacement of nearly pure near-surface ice—just as is indicated by the climate changes proposed by Head *et al.*¹.

Earth's present ice age, known as the Quaternary period, is only the latest and coldest phase of a progressive cooling over the past 43 million years. Earlier major ice ages occurred about 260 to 340 million years ago (during the Carboniferous and Permian periods), about 430 to 445 million years ago (during the Ordovician and Silurian periods), several times between about 520 and 950 million years ago (during the late Proterozoic era), about 2,250 million years ago (during the early Proterozoic era) and about 2,900 million years ago (during the Archeozoic era)⁷. As with Mars, the climatic changes of Earth's latest ice age had a pacemaker that operated through variation in the astronomical parameters of the planet's orbit and axis⁸. However, there is also a very strange element in this comparison. During its recent ice age, the martian poles had to warm up to release their store of frozen water into the atmosphere. For Mars, the episodes of ice



Figure 1 Ice age evidence. This image shows a striking, tongue-shaped feature, 4 km in length and an average of 800 m wide, taken by the Mars Orbiter Camera. The feature can be interpreted as a glacier of ice and rock flowing down the northern wall of a crater, 70 km in diameter, in the eastern Hellas region of Mars—and so of the occurrence of geologically recent ice ages on the planet.

accumulation ('glaciations') are warm phases for a planet that seems to be most stable in an ice-house condition. For Earth, the stable state seems to be the warm greenhouse condition, with cold phases corresponding to metastable periods of glaciation.

There is the controversial possibility that, for short periods, Earth might have become like Mars in this respect. This derives from the so-called 'snowball' conditions of the late Proterozoic, when glaciation reached Earth's equatorial latitudes⁹. How is it that such major instabilities are induced in the planet's climate? Mars also has very ancient landforms that are consistent with the incidence of epochs of past glaciation³. It

would seem that there are other causes that tip climatic stability from icehouse to greenhouse for Mars, and from greenhouse to icehouse for Earth. Although the orbital parameters can be the pacemakers for change within one of these states, they are not powerful enough to tip the balance between them.

The major causes of change could be volcanic or tectonic activity — which can, respectively, release radiatively active gases such as carbon dioxide, or alter the land and water configurations of the planetary surface, thereby altering heat distribution. Or they could be astronomical, for instance a change in the flux of interplanetary dust related to large-scale movements of the galaxy. With two planets on which to explore this puzzle, we should have a better chance of

resolving it than we would have by limiting our investigations to only one.

Victor R. Baker is in the Department of Hydrology and Water Resources, and the Lunar and Planetary Laboratory, University of Arizona, Tucson, Arizona 85721-0011, USA.

e-mail: baker@hwr.arizona.edu

1. Head, J. W., Mustard, J. F., Kreslavsky, M. A., Milliken, R. E. & Marchant, D. R. *Nature* **426**, 797–802 (2003).
2. Gilbert, G. K. *Am. J. Sci. (Ser. 3)* **31**, 284–299 (1886).
3. Baker, V. R. *Nature* **412**, 228–236 (2001).
4. Boynton, W. V. *et al. Science* **296**, 81–85 (2002).
5. Mellon, M. T. & Jakosky, B. M. *J. Geophys. Res.* **98**, 3345–3364 (1993).
6. Boynton, W. V. *et al. Eos* **84** (46), Fall Meet. Suppl. Abstr. P32B-05 (2003).
7. Crowell, J. C. *Mem. Geol. Soc. Am.* **192** (1999).
8. Hays, J. D., Imbrie, J. & Shackleton, N. J. *Science* **194**, 1121–1132 (1976).
9. Hoffman, P. F. & Schrag, D. P. *Terra Nova* **14**, 129–155 (2002).

Cell cycle

Passenger acrobatics

Toru Higuchi and Frank Uhlmann

Chromosomal passenger proteins undergo spectacular changes in localization during cell division. We now have molecular insight into how and why these changes occur.

Of all the cellular proteins, chromosomal passengers display some of the most visually striking patterns of behaviour. These proteins associate with chromosomes as cells prepare to segregate them during nuclear division (mitosis). Later, when chromosomes are aligned on the filaments that do the mechanical work of segregation, the passenger proteins become concentrated in distinct foci between paired kinetochores — the protein-based structures on chromosomes to which the filaments attach. Then, as chromosomes split, the passengers abruptly dissociate from the kinetochores and instead localize to the filaments in the area midway between the separating chromosomes¹ (Fig. 1a). Quite how these proteins relocate in this way has long been a mystery. Pereira and Schiebel² propose a solution in a paper in *Science*.

The first protein³ found to display such striking changes in localization during mitosis was named INCENP, for 'inner centromere protein' — the centromere being the DNA sequence onto which the kinetochore assembles. Two further proteins were later discovered to show a similar localization pattern: aurora B kinase, which regulates the activity of target proteins by transferring phosphate groups onto them, and a protein called survivin (reviewed in ref. 1). It has since become clear that these three proteins together form a stable aurora B kinase complex⁴. Earlier this year⁵, a fourth protein showing passenger behaviour was identified and named TD-60. This protein is required

to 'recruit' the aurora B kinase complex to kinetochores, and might regulate its activity.

So what does this kinase complex do? In the phases of the cell-division cycle that precede mitosis, the chromosomes are duplicated, producing pairs of 'sister chromatids'. The sisters must then be pulled apart to opposite poles of the cell by the mitotic spindle — an array of so-called microtubule filaments (see Fig. 1b). For this to happen successfully, sister kinetochores must attach to microtubules emanating from opposite poles of the spindle. The aurora B kinase complex is responsible for correcting the erroneous attachments that inevitably occur during the 'search-and-capture' mechanism that establishes bipolar attachment⁶. Later, as the chromosomes pull apart and aurora B kinase relocates to the region between them — the spindle's 'midzone' — it recruits microtubule-binding proteins. These stabilize the midzone, serving as a guide for the final step of cell division, the pinching off of the two daughter cells⁷.

The aurora B kinase complex, then, is clearly important at several stages of mitosis. But what lies behind its acrobatics? It was known that the timing of relocation coincides with the splitting of the chromosomes. So does relocation actually depend on this process? Apparently not. Chromosome segregation is triggered when the cohesive link between sister chromatids — a protein complex aptly named cohesin — is destroyed by the separase protein. But preventing chromosome segregation,

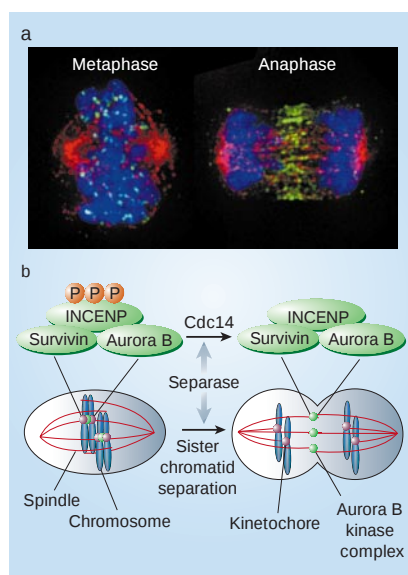


Figure 1 Moving passengers about.

Chromosomal passengers include the INCENP protein, and together they make up the aurora B kinase complex. a, During mitosis, this complex shows dramatic changes in localization, as these micrographs of dividing chicken DT40 cells reveal. In the metaphase of mitosis, INCENP (green) is found at kinetochores, the structures on chromosomes (blue) to which the mitotic spindle (red) attaches. As chromosomes split at the onset of anaphase, INCENP dissociates from kinetochores and moves to the midzone of the elongating spindle. b, Relocalization as explained by Pereira and Schiebel². It was known^{9,10} that, at the onset of anaphase, separase cleaves the cohesin protein (not shown) to separate sister chromatids, and at the same time activates the phosphatase Cdc14. Pereira and Schiebel find that Cdc14 removes phosphate (P) groups from INCENP, and that this triggers its relocalization to the spindle.

by making cohesin resistant to destruction by separase, does not stop aurora B kinase from relocating from the kinetochores to the spindle⁸.

Could passenger relocalization and chromosome segregation instead be caused by the same regulator? The answer is a clear 'yes' (Fig. 1b). Studies in budding yeast^{9,10} have shown that separase serves double duty: at the same time as separating sister chromatids, it also activates the protein phosphatase Cdc14. During the cell cycle, various proteins need to be phosphorylated so that cells can enter mitosis; once active, Cdc14 removes the phosphate groups, allowing cells to exit mitosis. Pereira and Schiebel² have now discovered that one of Cdc14's targets is the INCENP subunit of the budding yeast aurora B kinase complex. The authors show that INCENP dephosphorylation is both necessary and sufficient for the complex to relocate to the midzone of the mitotic spindle. The simple model, then, is

that separase activation separates sister chromatids and at the same time — via Cdc14 — triggers the relocalization of chromosomal passenger proteins.

This is a gratifying way to explain the striking coordination of chromosome segregation with aurora B kinase relocalization. It also opens up the possibility of addressing just how important the relocation of aurora B kinase is for its function. Pereira and Schiebel produced a version of INCENP that cannot be phosphorylated in mitosis, and so is located at the spindle even before chromosome segregation begins. One might have expected mitosis to be severely disrupted. But instead the effects were subtle: the chromosomes aligned on the spindle correctly, but the cells lost chromosomes more frequently. The implication is that the activity of the aurora B kinase complex is not determined solely by its location. It would be interesting to complement this study by analysing an INCENP mutant that mimics permanent phosphorylation. Would this prevent it from leaving kinetochores? If so, what would the consequences be?

When chromosomes segregate in the absence of Cdc14 activity, the spindle is grossly disorganized and breaks down prematurely^{11,12}. Knowing that INCENP is a target of Cdc14, Pereira and Schiebel also investigated whether the non-phosphorylatable INCENP could compensate for the absence of Cdc14 in stabilizing the spindle. They found that the mutant INCENP recruited midzone proteins to spindles in some cells — but the spindles still broke.

Evidently, then, Cdc14 has targets in addition to INCENP that coordinate the assembly of intact spindles.

Next to nothing is known about how Cdc14 is regulated in organisms other than yeast. INCENP is phosphorylated prior to mitosis in vertebrates, too⁴, so it is quite possible that separase induces Cdc14 to trigger passenger relocalization in higher organisms as well. By revealing how the aurora B kinase complex relocalizes, Pereira and Schiebel² have explained a beautiful aspect of cell biology. The more challenging task that remains is to discover what exactly the complex does at kinetochores and at the spindle midzone, to bring about faithful completion of cell division. ■

Toru Higuchi and Frank Uhlmann are in the Chromosome Segregation Laboratory, Cancer Research UK London Research Institute, 44 Lincoln's Inn Fields, London WC2A 3PX, UK. e-mail: frank.uhlmann@cancer.org.uk

1. Adams, R. R., Carmona, M. & Earnshaw, W. C. *Trends Cell Biol.* **11**, 49–54 (2001).
2. Pereira, G. & Schiebel, E. *Science* published online 6 November 2003; doi:10.1126/science.1091936.
3. Cooke, C. A., Heck, M. M. & Earnshaw, W. C. *J. Cell Biol.* **105**, 2053–2067 (1987).
4. Bolton, M. A. *et al. Mol. Biol. Cell* **13**, 3064–3077 (2002).
5. Mollinari, C. *et al. Dev. Cell* **5**, 295–307 (2003).
6. Tanaka, T. U. *et al. Cell* **108**, 317–329 (2002).
7. Severson, A. F., Hamill, D. R., Carter, J. C., Schumacher, J. & Bowerman, B. *Curr. Biol.* **10**, 1162–1171 (2000).
8. Hauf, S., Waizenegger, I. C. & Peters, J.-M. *Science* **293**, 1320–1323 (2001).
9. Stegmeier, F., Visintin, R. & Amon, A. *Cell* **108**, 207–220 (2002).
10. Sullivan, M. & Uhlmann, F. *Nature Cell Biol.* **5**, 249–254 (2003).
11. Sullivan, M., Lehane, C. & Uhlmann, F. *Nature Cell Biol.* **3**, 771–777 (2001).
12. Gruneberg, U., Glotzer, M., Gartner, A. & Nigg, E. A. *J. Cell Biol.* **158**, 901–914 (2002).

Astronomy

Wide-angle lens

Joachim Wambsganss

Gravitational lenses produce multiple images of single astronomical objects. The most widely separated images of a quasar ever found reveal the dark-matter content of the lensing galaxies.

Record-breaking is usually gently incremental, each record a slight improvement on the last. It's rare that a new record more than doubles the old record-holding value, particularly one that had stood for more than two decades. But such a case is reported on page 810 of this issue: Naoshi Inada and colleagues¹ have discovered multiple images of a quasar, created by gravitational lensing, whose maximum angular separation across the sky is 14.62 arcseconds. The previous record for a quasar stood at 6.3 arcseconds.

Gravitational lensing is a consequence of Einstein's general theory of relativity. If a massive object such as a star, a galaxy or a galaxy cluster happens to be close to the line of sight from Earth to a more distant source,

the gravitational attraction of this 'lens' deflects the light from the distant source. To the observer on Earth, the source may appear magnified, distorted and — most dramatically of all — as multiple images on the sky.

Einstein himself was sceptical that this effect would ever be seen — "There is no great chance of observing this phenomenon", he said². But so far, around 80 multiply-imaged quasars have been found³. A quasar is the very luminous central part of an active galaxy, powered by a super-massive black hole. The first lensed quasar⁴, Q0957+561, or 'Old Faithful', was discovered in 1979. The two images created by the intervening lens are separated by 6.3 arcseconds (an arcsecond is 1/3,600 of a degree), and Old Faithful has held the record for the



100 YEARS AGO

The adequate provision of secondary and higher education for English girls and women is to be regarded as one of the accomplishments of the latter half of the nineteenth century. In 1850, for instance, the popular idea here and elsewhere was that women were intellectually incapable of benefiting by higher instruction. To quote Dr. Leslie Waggener, of the University of Texas, "it was seriously questioned whether the 'female' mind could untangle the intricacies of pure mathematics, could appreciate the abstruse speculations of metaphysics, or could follow, step by step, the inductions of a scientific investigation." Fifty years' experience has, however, demonstrated the complete fallacy of this preconception... So complete a change of opinion on a subject of such importance as the suitable education of the larger half of the human race deserves attention, and the steps in the movement which has resulted in the recognition of the claims of women at most universities throughout the world, supply a profitable study for all students of educational problems. A comparison, too, of the present provision of university courses for women with their complete non-existence in 1850 should serve to cheer those men of science and others who are endeavouring to improve our national education in other directions.

From *Nature* 24 December 1903.

50 YEARS AGO

In his speech to the General Assembly of the United Nations at New York on December 8, President Eisenhower said that the United States are prepared immediately to meet privately with such other nations as may be "principally involved" to seek "an acceptable solution" to the atomic armaments race... The President indicated that he is prepared to submit to the United States Congress, with every expectation of approval, plans which would encourage world-wide investigation into the most effective peace-time uses of fissionable material and which would begin to diminish the potential destructive power of the world's atomic stock-piles. Such plans should also demonstrate to peoples of all nations that the great Powers of the world, both of the East and of the West, are interested in human aspirations first and foremost, rather than in building up armaments of war.

From *Nature* 19 December 1953.

largest known splitting of quasar images ever since. (In fact, another candidate⁵ exists with a separation of 6.9 arcseconds, but its identification as a lensed quasar is not 100% certain yet.) Of the other quasar lenses that have been discovered over the years, most are double or quadruple images. Some of the images are circular, and are called Einstein rings⁶.

Spotting a quasar that is gravitationally lensed among all the objects in the sky is not at all easy. Astronomers must establish that the suspected multiple images are all at the same distance from Earth, and that their spectra of emitted radiation — as characteristic of an astronomical object as a genetic fingerprint is of a person — are identical, or at least very similar. Ideally, the gravitational lens itself (which is typically a massive galaxy) should be visible between the quasar images, although this may not always be so. But the ultimate proof of the lensing effect hangs on the variable nature of quasars: does the apparent brightness of each image vary in the same way, allowing for a certain time delay between them⁷?

The evidence is strong that the four images recorded by Inada *et al.*¹ correspond to a single gravitationally lensed quasar, rather than a group of quasars. The authors have shown convincingly that their lensed quasar fulfils the first three of the above criteria. The variability requirement is still to be met, but this can easily take a few years of observation. The time delay in brightness variations between the multiple images arises because the signals travel along different geometrical paths. Combined with the time dilation that is predicted by general relativity in strong gravitational fields, the journey times along the different paths may be different by months, or even years. And, of course, observers need a bit of luck too: the quasar itself must cooperate and produce some brightness fluctuations during the period of observation.

But what is so interesting about widely spaced multiple images of a lensed quasar, anyway? This quasar (SDSS J1004+4112)¹ is more than a mere record-breaker. Although 'weighing' cosmic objects is generally a very difficult task in astronomy, here the splitting angle between the images is a clean measure of the mass of the lens. The mass of this lens is much higher than that of typical galaxies. Inada *et al.* have identified the lens as a bright galaxy, together with a cluster of surrounding galaxies at the same distance from Earth. But the visible mass in these galaxies falls short, by a large factor, of the amount needed to explain the wide splitting of the images. The extra mass must be dark matter.

Most of the known Universe is made of dark matter, and galaxies are expected to be surrounded by haloes of it. For galaxies in clusters, these dark-matter haloes have been measured so far only through weak gravitational-lensing effects⁸. But the lensing effect

seen by Inada *et al.*¹ is very strong, and this is the first detection of a galaxy cluster whose very massive dark halo is splitting a quasar image. The fact that they see four images also hints at the mass profile: the dark matter must be quite concentrated, and slightly asymmetric, around the cluster. Once the variability of the quasar has been picked up in each of the images, that will reveal⁹ more about the distribution of dark matter in this system. But, as the authors state, it is already clear that this discovery confirms the predictions of the 'concordance' model⁹, the standard model of cosmology.

Inada *et al.*¹ found this record-breaking quasar lens by performing a systematic search among the tens of thousands of quasars logged by the Sloan Digital Sky Survey (SDSS)¹⁰. Previously, the most comprehensive lens searches had been done at radio wavelengths¹¹; they showed that about one in 700 quasars is multiply imaged by a gravitational lens. But with the advent of two large-scale surveys, SDSS and 2dF (ref. 12), such huge studies are now possible at

visible wavelengths as well. Many more large-separation quasar lenses should soon be found, enabling quantitative comparison of the matter distribution in the Universe with theories of its formation. It's a safe bet that it won't be another 24 years until this quasar record is broken.

Joachim Wambsgans is at the Physics Institute, University of Potsdam, Am Neuen Palais 10, 14469 Potsdam, Germany.

e-mail: jkw@astro.physik.uni-potsdam.de

1. Inada, N. *et al.* *Nature* **426**, 810–812 (2003).
2. Einstein, A. *Science* **84**, 506 (1936).
3. <http://cfa-www.harvard.edu/castles>
4. Walsh, D., Carswell, R. F. & Weymann, R. J. *Nature* **279**, 381–384 (1979).
5. Muñoz, J. A. *et al.* *Astrophys. J.* **546**, 769–774 (2001).
6. Wambsgans, J. *Living Reviews in Relativity* **1**, No. 12 (1998); www.livingreviews.org/Articles/Volume1/1998-12wamb
7. Natarajan, P. & Kneib, J.-P. *Mon. Not. R. Astron. Soc.* **287**, 833–847 (1997).
8. Kochanek, C. S. *Astrophys. J.* **578**, 25–32 (2002).
9. Ostriker, J. P. & Steinhardt, P. *Nature* **377**, 600–602 (1995).
10. Richards, G. T. *et al.* *Astron. J.* **123**, 2945–2975 (2002).
11. Browne, I. W. A. *et al.* *Mon. Not. R. Astron. Soc.* **341**, 13–32 (2003).
12. Croom, S. M. *et al.* *Mon. Not. R. Astron. Soc.* **325**, 483–496 (2001).

Ecology

Badger cull culled

Timothy J. Roper

Large-scale field trials have been under way to assess how effective badger culls are in stemming the incidence of tuberculosis in cattle. One culling tactic, it seems, increases occurrence of the disease.

If a disease is transmitted to humans or domestic animals by wildlife, an obvious policy option is to cull the wildlife that harbour the disease. This has been the logic underlying attempts in Britain to control bovine tuberculosis (caused by *Mycobacterium bovis*) in cattle, through the culling of badgers.

Over the past few years, trials have been carried out to assess the effectiveness of two different badger-culling strategies. But last month the alarming initial results from one of the trial 'treatments', known as reactive culling, prompted the UK government to call a halt to this aspect of the trials. Given the urgency of the situation, the results were first released at a press conference, then in an online publication. They now appear in print (see the paper by Donnelly *et al.*¹ on page 834).

The transmission of bovine tuberculosis from wildlife to livestock is not a uniquely British problem. It is equally serious in Ireland, has recently been reported in Michigan in the United States, and has also occurred in Australia and New Zealand. Although the disease can be transmitted to humans, the main reason for concern is economic, because of its impact on beef and dairy farmers, rather than health-related. In Britain, a

government agency — the Department for Environment, Food and Rural Affairs — spends about £74 million (US\$125 million) annually trying to control the disease, with further costs to the farming industry. These costs would escalate steeply if the rate of infection of cattle herds continues to increase, as it is now doing exponentially.

Bovine tuberculosis is especially difficult to deal with in Britain because of its incidence in badgers (Fig. 1), which are a protected species and are now a large and widespread source of infection. Indeed, the high population density of badgers in Britain and Ireland probably explains why bovine tuberculosis is more prevalent in these countries than in continental Europe. Hence the policy of culling badgers. Culling of a wildlife reservoir of tuberculosis does sometimes work. In New Zealand, control of brushtail possums resulted in a 50% reduction in the number of infected cattle and deer herds between 1994 and 2000. In Australia, culling of feral water buffalo has contributed to eradication of the disease².

The results of Donnelly *et al.*¹, however, suggest that localized culling of badgers in Britain has actually increased the incidence of bovine tuberculosis in cattle. This result demolishes the case for localized culling as



Figure 1 Young badgers and (inset) a transmission electron micrograph of *Mycobacterium bovis*, the bacterium responsible for bovine tuberculosis.

a method of controlling the disease, while paradoxically providing further evidence that badgers are somehow implicated in its epidemiology. The difficulty is that there is no obvious alternative method of controlling either badgers or bovine tuberculosis.

Culling of badgers in tuberculosis 'hotspots' in Britain has been taking place since the mid-1970s — despite its unpopularity with farmers, who have complained about its ineffectiveness, and with badger conservationists, who have disputed whether badgers transmit the disease to cattle. In an attempt to test the effectiveness of the strategy once and for all, in 1997 a scientific committee³ recommended that a randomized culling trial should be carried out in parts of Britain suffering from a high incidence of cattle tuberculosis. The trial was intended to involve ten replicates of three different treatments: proactive culling, where as many badgers as possible were culled within each treatment area; reactive culling, where localized culling was carried out only when cattle in the relevant area had tested positive for bovine tuberculosis; and no culling, where no action was taken to reduce badger populations.

Donnelly *et al.* report the results from the trial designed to test the effectiveness of reactive culling. (The proactive-cull treatment will continue, but those results are not expected for some time.) Comparison of the incidence of bovine tuberculosis in cattle in nine reactive and ten no-culling areas reveals a 27% increase in cattle tuberculosis following culling, with a higher than expected incidence of the disease in all nine reactive areas. Given the practical difficulties of carrying out such an extensive field trial — including

non-compliance by farmers, disruption by animal-rights activists, and suspension of the trials during an outbreak of foot-and-mouth disease — this is a remarkably large and consistent effect. It has also appeared surprisingly rapidly, in that reactive culling commenced less than 18 months ago in most of the areas concerned.

What could be the mechanism underlying the increased incidence of disease? There are no straightforward answers. Ecologists, including the UK government's own scientific advisers, have long suspected that culling could disrupt the otherwise stable social-territorial system of badgers, resulting in wider dispersal and ranging behaviour⁴. As Donnelly *et al.* acknowledge, this does indeed happen around the margins of culled areas⁵. But for badger movements to increase the incidence of tuberculosis in cattle within the culled area, either more infected badgers would have to migrate into it than were removed in the cull, or they or any remnants of the original population would have to behave differently, in a way that would substantially increase the probability of disease transmission to cattle. The first possibility seems unlikely, because previous studies⁴ suggest that recolonization of culled areas proceeds slowly; the second is pure speculation.

Alternatively, perhaps the absence of badgers results in an increase in some other species that transmits tuberculosis to cattle. But bovine tuberculosis only infects mammals, and badgers prey mainly on invertebrates, so this explanation seems implausible. It also fails to explain two other points: the immediacy of the effect (the relevant species would have to increase in

numbers, become infected, develop the disease to the stage at which it can be passed on, and infect cattle which must then be tested); and why previous culls, in other locations, have produced decreases in cattle tuberculosis^{2,3}. All in all, this is a sobering reminder of how little is understood about wildlife population dynamics and about the epidemiology of bovine tuberculosis.

Regardless of the mechanism underlying the increase in disease associated with reactive culling, the political consequences are unambiguous: this part of the trial has been terminated and it is unlikely that future strategies for controlling bovine tuberculosis will include badger culling. In the short term, attention is likely to switch to control methods involving factors related to cattle husbandry and farm management, such as keeping badgers away from cattle and preventing them from contaminating stores of cattle food⁶. Such measures have not yet been tested, however, and it is impossible to predict their effectiveness, because it is not yet known how badgers transmit tuberculosis to cattle.

In the longer term, Donnelly and colleagues' results are likely to strengthen calls for the development of a vaccine for cattle or badgers, or both². As regards cattle, vaccination is effectively prohibited under current European Union legislation, because it would compromise the skin test used to diagnose tuberculosis infection. So it will be necessary to develop not just a suitable vaccine, but also a diagnostic test that will distinguish between vaccinated cattle and infected ones³.

With badgers, the major hurdle is the practical one of delivering immunization to a large enough proportion of the host population². Attention has focused mainly on bait as a vehicle for vaccine delivery, not least because this method has proved an effective method of rabies control in foxes and jackals². New vaccines, possibly based on the inactivation of virulence genes in *Mycobacterium bovis*, and methods of delivering them to wildlife, are being investigated in Britain, the Republic of Ireland and New Zealand. But the scientific and practical problems are considerable — solving them could take a decade or more^{3,7}.

Timothy J. Roper is in the Department of Biology and Environmental Science, University of Sussex, Brighton BN1 9QG, UK.

e-mail: T.J.Roper@sussex.ac.uk

1. Donnelly, C. A. *et al.* *Nature* **426**, 834–837 (2003).
2. de Lisle, G. W. *et al.* *Rev. Sci. Tech. Off. Int. Epiz.* **86**–111 (2001).
3. Krebs, J. *et al.* *Bovine Tuberculosis in Cattle and Badgers* (PB3423, MAFF, London, 1997).
4. Cheeseman, C. L. *et al.* in *The Badger* (ed. Hayden, T. J.) 78–93 (Royal Irish Academy, Dublin, 1993).
5. Tuijthens, F. A. M. *et al.* *J. Anim. Ecol.* **69**, 815–828 (2000).
6. Garnett, B. T. *et al.* *Proc. R. Soc. Lond. B* **269**, 1487–1491 (2002).
7. Skinner, M. A. *et al.* *Rev. Sci. Tech. Off. Int. Epiz.* 112–132 (2001).

This article was published online on 23 November 2003 (doi:10.1038/nature02219).



Plant physiology

Needles point to pollution

Plant Cell Environ. **26**, 1929–1939 (2003)

The iconic shape of conifers — a single vertical main trunk with minor side branches — is distorted by exposure to air pollution. According to M. D. Collier and colleagues, the hormonal changes involved could provide an early warning of nitrogenous pollutants.

Conifer growth is an example of ‘apical dominance’, where a plant’s growing tip prevents the development of major side branches. Gardeners make trees and shrubs more bushy by suppressing apical dominance through pruning; plants achieve the same effect themselves by using hormones called cytokinins.

Cytokinins have previously been linked to loss of apical dominance in spruce trees growing in polluted areas. To investigate this further, Collier *et al.* sprayed plantations of Sitka spruce with combinations of ammonium nitrate and sodium sulphate. Although the three-year study period was too short to show changes in growth patterns, the cytokinin content of the trees’ needles was substantially increased after exposure to the nitrogenous pollutant.

These increases were masked by simultaneous spraying with acid sulphur. But this may not bar cytokinin from becoming a biological indicator of nitrogen

pollution. In Europe at least, industrial sulphur emissions are dropping, whereas nitrogenous pollution is increasing. Such trends will need careful monitoring if the Christmas tree is not to become the Christmas bush.

Christopher Surridge

Neurobiology

Amnesia in ageing flies

Neuron **40**, 1003–1011 (2003)

Like humans, fruitflies gradually lose their memory as they age. Takuya Tamura and colleagues show that ageing specifically impairs the fly’s middle-term memory — a phase of memory formation that depends on the *amnesiac* gene.

Flies can learn to avoid odours when specific fragrances are coupled with electric shocks. Their memories can be separated into three distinct phases: short-term memory is present immediately after training, and fades away within 2 hours; middle-term memory lasts more than 5 hours; and an ‘anaesthesia-resistant’ memory lasts more than 24 hours. These three phases seem to form independently to some extent, as mutations in ‘memory’ genes affect each phase differently. Mutations in *amnesiac*, for instance, disrupt middle-term memory only.

Tamura and colleagues find that the memory defects of older flies are identical to those of flies with *amnesiac* mutations. Also, the mutant flies do not show any further memory decline as they age. Restoring *amnesiac* expression in the mutants brings back memories of smells and shocks — but also reinstates age-dependent memory decline.

Yet expression of *amnesiac* is not altered in normal old flies, nor can their memory be repaired by more copies of the gene. The authors propose that although age-related memory decline may result from disruption of the *amnesiac* pathway, this gene itself seems unaffected.

Marie-Thérèse Heemels

Nonlinear dynamics

Chaos relatively robust

Phys. Rev. Lett. **91**, 231101 (2003)

The behaviour of chaotic systems is acutely sensitive to their initial conditions and to the way the trajectories of their components evolve in time. So it wouldn’t seem surprising if a change in coordinate systems — the kind of transformation invoked in general relativity — were to perturb this behaviour. On the contrary, Adilson E. Motter finds that chaos seems in general to be indifferent to changes in coordinates. It is, in other words, relativistically invariant.

Chaos can be quantified by the Lyapunov exponent, a measure of how quickly small initial differences diverge

as two near-equivalent systems evolve. A positive Lyapunov exponent is a signature of chaos. Motter finds that a change in time coordinates may alter the absolute value of Lyapunov exponents but does not alter their sign: chaos cannot be introduced or conjured away by a relativistic reparametrization of time.

This conflicts with an earlier suggestion that at least one model of quantum cosmology (an attempt to match quantum theory with relativity) can have either a positive or zero Lyapunov exponent, depending on the choice of coordinates. That outcome, says Motter, arises from an improper definition of the Lyapunov exponent.

Phillip Ball

Meteors

Parallel lines

Geophys. Res. Lett.

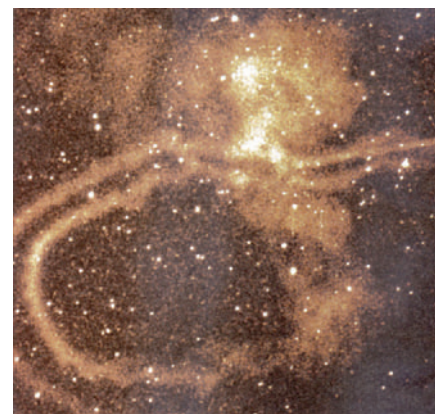
doi:10.1029/2003GL018312 (2003)

Meteors passing through the upper atmosphere can leave behind them glowing trails, like the contrails of aircraft. But this signature is sometimes written with a double pen: a single meteor leaves two parallel trails, which are subsequently twisted and distorted by winds. What causes these ‘double trains’?

A long-standing idea is that double trains — recognized in meteor showers such as the Leonids since at least the beginning of last century — are the two sides of a hollow ‘tube’ ploughed out by the meteor, the darker interior of which is created by chemical depletion of the light-emitting species. But the brightness contrast is too big for this to be tenable.

Perhaps the meteor fragments into two? But if so, why never into three or four? M. C. Kelley *et al.* argue that, instead, the shape and dynamics of double trains are consistent with the idea that one train is due to excited gas from the meteor, and the other, lying below the first, comes from dust sedimenting under gravity.

Phillip Ball



Efficiency of equine express postal systems

Relay riders over two millennia delivered mail with a remarkably consistent alacrity.

Express postal systems relied on horses for a period that spanned two millennia (540 BC to AD 1861), and these systems showed remarkable consistency in the average mail-courier speed and in the distance between horse-changing stations. Here I show how this adopted speed and distance combination was ideal for optimal horse performance and can be explained with a modern understanding of equine physiology and with reference to recent endurance records. The parameters of the historical systems were chosen to avoid heat stress deriving from an otherwise over-boosted metabolic machine and to reduce the risk of the animal falling lame.

The design of ancient, horse-driven postal systems may have empirically converged to an optimized exploitation of horse and rider, compatible with their physiological and biomechanical burdens. For example, Xenophon¹ reported that before constructing the Royal Persian Road (Fig. 1), a route of 2,750 km that was travelled in seven days and nights by fast-relay couriers, King Cyrus the Great (599–530 BC) “experimented to find out how great a distance a horse could cover in a day when ridden hard, but so as not to break down, and then he erected post-stations at just such distances”.

The most recent and well-documented horse-based mail system is the Overland Pony Express, which operated during 1860–61 in the United States between Sacramento, California, and St Joseph, Missouri, for the benefit of west-coast pioneers.

Table 1 (overleaf) shows results from a historical analysis of express postal systems (ref. 2, and see supplementary information). Despite differences in latitude and civilization, some invariant features of such systems emerge: a distance between horse-changing stations of about 20–25 km, 4–6 stations travelled by a single rider, and an average speed of about 16 km h⁻¹. The riders' endurance capability was limited by their capacity for aerobic metabolism: the metabolic cost of horse-riding is 1.9–2.4 litres of oxygen per minute (see supplementary information), which corresponds to 66.4% of maximum metabolic power and can be sustained for up to 276 min — equivalent to 92 km, or 4–5 stations.

The average speed of the courier, which corresponds to the transition between trot and gallop by the horse^{3,4}, was probably determined by a faster rate of travel (say, 20–22 km h⁻¹) during the day and a slower one (about 10–12 km h⁻¹) at night. These speeds correspond to the optimum (that is, the minimum



Right, the world's best performances by horses in short-distance (triangles; www.horseworlddata.com/qhwrec.html), middle-distance (light grey circles; www.horsemanpro.com/articles/gallop.htm) and endurance (dark grey circles; www.ansi.okstate.edu/breeds/horses) races. Data for human running performance (open circles) are shown for comparison. Logarithmic regressions were carried out separately on the middle-distance and endurance results. The shaded region corresponds to the range of distances between horse-changing stations on postal express routes. 'Turbo' is the power boost that is generated by both anaerobic metabolism and emptying of the equine spleen.

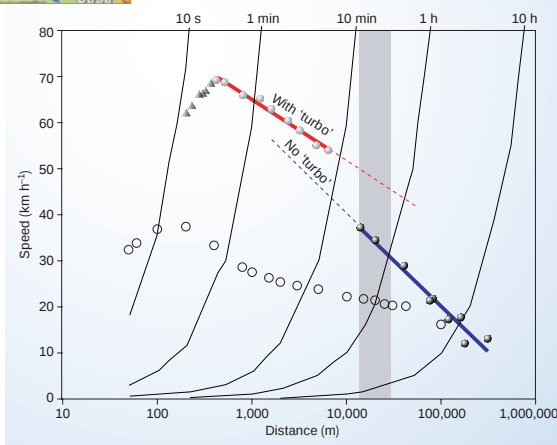


Figure 1 Express postal systems and horse performance. Left, map of the Persian Royal Road (540 bc; shown in red), the oldest horse mail route, organized by King Cyrus the Great. The invariant courier speed and distance between horse-changing stations on this road, as well as in other express postal systems of the past, bear witness to an awareness of the limitations of horse performance over long distances.

metabolic cost per unit of distance) speed of galloping and trotting, respectively^{3,4}.

Horses aged 4–7 years, with a mass of 390 kg (see www.xphomestation.com/facts.html), have a metabolic power of 15–28 l O₂ min⁻¹ when galloping at 20 km h⁻¹ (refs 3–5), which corresponds at worst to ‘just’ 69% of their maximum aerobic power (measured⁶ as 40.5 l O₂ min⁻¹). It is therefore surprising that horses needed to be replaced every 1.5 hours, whereas less athletic species, such as humans, can exercise at the same metabolic load for more than 4 hours (ref. 7).

To investigate this discrepancy, I analysed actual speed records for horse-racing and endurance riding (Fig. 1). When plotted against the logarithm of the race distance, a sudden drop in performance is apparent for middle-to-long distances, although the two data sets are individually well fitted by separate (logarithm) functions. Unlike humans running over a range of distances, where there is a relatively smooth decrease in the curve, due mainly to the progressive depletion of glycogen stores⁷, horses experience a power boost for races of short-to-middle distance and a decline in performance at distances greater than 10 km.

The factors that determine this performance over different distances, and which therefore influence the strategy adopted by

historical postal systems, can be identified from our knowledge of equine exercise physiology. According to the intensity of exercise, the sympathetic nervous system⁸ controls the emptying of the spleen, which contains 30% of the total number of red blood cells, into the bloodstream. This internal blood ‘doping’ causes the haematocrit to increase from 40% to 65–70% and, in conjunction with anaerobic metabolism, is important for high performance in competitions over short-to-middle distances.

Such ‘turbo-charging’ of the metabolic machinery leaves horses vulnerable to cooling problems, however, during long-lasting events. An animal with muscle as 55–60% of its body mass generates a huge amount of heat (equal to three times the mechanical work done⁹), which must be dissipated as sweat (up to 10–15 litres h⁻¹). This burden reduces the sustainable cruising speed and renders the ‘boost’ option useless during endurance rides.

Over long distances, a horse's adrenergic threshold is 32 km h⁻¹ (ref. 10) and its haematocrit is about 46–49% (ref. 11; this increase is partly due to blood concentration), meaning that the risk of heat stress/stroke and pulmonary hypertension/haemorrhage (contributed by plasma loss and blood hyperviscosity) is reduced¹². On the other hand, very long journeys

Table 1 Features of some horse-driven postal systems

Date	Mail system	Route	Distance covered		Average speed (km h ⁻¹)	Number of stations	Network extent (km)
			Horse (km)	Rider (stations)			
540 BC	Cyrus the Great, Persian Empire	Sardis–Susa	24.3	–	15.3	111	2,757
AD 618	Tang Dynasty, China	Network in China	20.0	5–6	13.3	1,297	32,500
AD 1250	Kublai Khan, Mongol Empire	Network in Asia	18.0–25.0	–	15.0–20.4	1,400	60,000
AD 1260	Mamelouk Sultanate, Egypt, Syria*	Cairo–Damas	20.0–25.0	–	16.1	200	3,000
AD 1425	G. G. Visconti, Torre e Tasso, Italy	Network in Europe	16.1	–	14.8	–	–
AD 1477	Louis XI, France	Network in France	28.0 ± 1.2 s.d.	–	13.8–16.5	72	2,000
AD 1860	Overland Pony Express, USA	St Joseph–Sacramento	20.1	4–5	15.5	157	3,163

Although express postal systems date back to 1780 BC (Hammurabi), only those based on day-and-night express relay riding, for which reliable data and historical sources (ref. 2, and see supplementary information) are available, have been included. The Roman Cursus Publicus (AD 1), despite its huge extent, was not considered here because of its slowness¹³.
*Both horses and camels (*Camelus dromedarius*) were used in this system.

may cause neuromuscular fatigue, musculo-skeletal damage, altered kinematics of gallop, and lameness¹².

On this basis, it seems that relay postal systems selected the greatest speed and distance that were compatible with a low risk of damage to the horse (no spleen emptying, no anaerobic metabolism, adequate cooling; Fig. 1b, lower line), which was a valuable investment. The speed of 20–22 km h⁻¹ used for delivery by the postal systems in daylight hours was slower than the speed of 35 km h⁻¹ shown in Fig. 1 for elite, modern horses that have been subject to superior selection, training, feeding and resting regimes. Moreover, inclines and obstacles would have reduced the average speed of a postal horse.

It is remarkable how several civilizations in the history of postal systems, without any

knowledge of equine physiology, independently worked out the same optimal parameters (maximum distance travelled and related speed) for reducing the risk of life-threatening conditions such as heat stress and lameness in their horses. The distance between the postal stations of the past finds a parallel in the distance between veterinary checkpoints in modern endurance races, in which cardiovascular, metabolic and locomotory status are monitored every 20–30 km. Only the rider who simultaneously proves to be the fastest and finishes the race with the horse in good condition is eligible as the winner.

Alberto E. Minetti
Institute for Biophysical and Clinical Research into Human Movement, Manchester Metropolitan University Cheshire, Alsager ST7 2HL, UK
e-mail: a.e.minetti@mmu.ac.uk

1. Xenophon *Cyropaedia* Book VIII, 6.17–18.
2. Gazagnadou, D. *La Poste à Relais* (Kimé, Paris, 1994).
3. Hoyt, D. F. & Taylor, C. R. *Nature* **292**, 239–240 (1981).
4. Minetti, A. E., Ardigo, L. P., Reinach, E. & Saibene, F. *J. Exp. Biol.* **202**, 2329–2338 (1999).
5. Hanak, J. et al. *Acta Vet. Brno* **70**, 133–139 (2001).
6. Potard, U. S., Leith, D. E. & Fedde, M. R. *J. Appl. Physiol.* **84**, 2052–2059 (1998).
7. Saltin, B. in *Limiting Factors of Physical Performance* (eds Keul, J. & Thime, G.) 235–252 (Georg Thieme, Stuttgart, 1973).
8. Snow, D. H. *Res. Vet. Sci.* **27**, 372–378 (1979).
9. Woledge, R. C., Curtin, N. A. & Homsher, E. *Energetic Aspects of Muscle Contraction* (Academic, London, 1985).
10. Snow, D. H., Harris, R. C., MacDonald, I. A., Forster, C. D. & Marlin, D. J. *Equine Vet. J.* **24**, 462–467 (1992).
11. Rose, R. J. *Br. Vet. J.* **142**, 542–552 (1986).
12. Foreman, J. H. *Vet. Clin. North Am. Equine Pract.* **14**, 205–219 (1998).
13. Suetonius *The Twelve Caesars* Book I, 57.

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Structural colour

Opal analogue discovered in a weevil

Beetles in dimly lit tropical forests often display structural colours¹, but in direct sunlight only part of the insect can be seen from any direction — it appears as a spot of light because multilayer reflectors on its rounded surface act like mirrors^{2,3}. Here we describe a beetle, *Pachyrhynchus argus*, found in forests in northeastern Queensland, Australia, that has a metallic coloration that is visible from any direction owing to a photonic crystal structure analogous to that of opal⁴. To our knowledge, this is the first recorded example of an opal-type structure in an animal.

The weevil *P. argus* has a relatively uniform, metallic colour (Fig. 1a), even though the light in its environment can be strongly directional. This colour derives from scales that are about 0.1 mm in diameter (Fig. 1b) and occur in patches on the top and sides of the beetle's roughly hemispherical body. Individually, the scales are flat, lying parallel with the body, and consist of two parts: an outer shell and an inner structure.

The inner structure is absent in some scales, causing the complete scale to appear transparent. Otherwise, in transmitted white

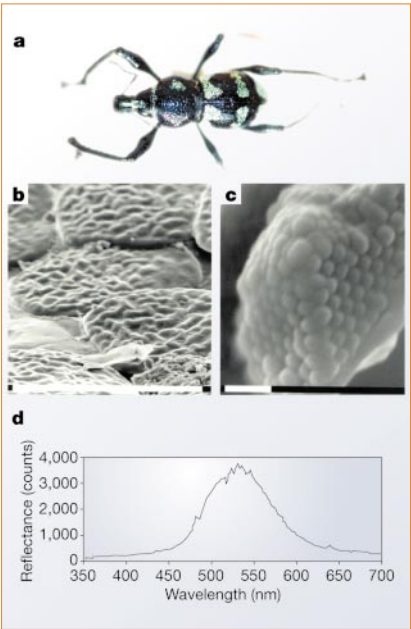


Figure 1 Structure and optical properties of the scales of the beetle *Pachyrhynchus argus*. **a**, Dorsal view. The beetle's metallic coloration is visible from every direction owing to the optical properties of its scales, which act as a photonic crystal and have a structure analogous to that of opal. **b**, Scanning electron micrograph (SEM) of several partially overlapping scales (green in **a**). **c**, SEM of a section through a single scale, showing the 'opal' structure fractured (as an artefact of sectioning) to form a pyramid shape. **d**, Spectrometric analysis. White light was incident at 20° to the normal of the scale surface, and the reflectance profile was measured at 20° to the other side of the normal. Scale bars: **b**, 100 μm; **c**, 1 μm.

530 nm, at an angle of 20° to the other side of the normal (Fig. 1d).

We examined sectioned scales internally using scanning and transmission electron microscopes. The inner structure of the scales is a solid array of transparent spheres, each with a diameter, *d*, of 250 nm. They are arranged in flat layers and have a precise, hexagonal close-packing order (Fig. 1c). As the lattice constant of the crystal approaches half the wavelength of light, it acts as a three-dimensional diffraction grating⁵, forming optical domains within a single scale. This is the most common nanostructure found in opal (although opal microspheres occasionally show face-centred cubic packing)^{6,7}, and allows for the reflection of a narrow range of wavelengths over a wide range of incident angles.

This opal-type structure is an example of a three-dimensional 'photonic crystal'. Using the formula $\lambda_{\max} = 2d \times 0.816 \sqrt{(n^2 - \sin^2 \theta)}$ as an approximation (for details, see www.icmm.csic.es/cefe/Infiltration/R6G/R6G_infill.htm), the wavelength of maximum reflectance, $\lambda_{\max}$, for an angle of incidence θ , was calculated as 573 nm at $\theta = 20^\circ$. The constant $0.816 = 2/3^{1/2}$ accounts for the spacing between the close-packed planes in units of sphere diameter; n represents the average refractive index in the system where point scatterers are arranged in planes, calculated as $(n_s + n_m)/2$, where n_s is the refractive index of the microspheres and is taken as 1.56, and n_m is the refractive index of the matrix (1.33) — that is, the refractive index of the chitinous material and 'water' that make up the beetle's exoskeleton^{7,8}, respectively.

This is a good match with the measured $\lambda_{\max}$ and indicates that the cause of the optical effect is indeed similar to that of opal⁷. Variations in θ do, however, cause differences in $\lambda_{\max}$: the range of θ values of $0-70^\circ$ equates to a range of $\lambda_{\max}$ from 590 nm to 448 nm. The invariant colour (yellow-green) seen from the whole animal is a result of global averaging of the different domains within each scale and juxtaposed scales; the domain structure thus creates omnidirectional colour, removing the iridescent effect.

The first photonic crystal revealed as such in an animal (in this case a polychaete)⁹ has a two-dimensional structure. By contrast, the three-dimensional properties described here allow for a relatively omnidirectional optical effect, which is important to the behaviour of this weevil because it appears to be strongly coloured from every direction *in situ*. This could be useful for interspecific colour or pattern recognition.

Opal is notoriously difficult to manufacture in solid form. However, transmission electron micrographs of the weevil's opal-like structure reveal repeating patterns of light and dark areas within each microsphere, providing a clue to their molecular structure (which may be revealed by X-ray diffraction analysis) and production. These microspheres must be constructed by molecular self-assembly, a process that could potentially be reproduced¹⁰ by the synthetic-opal industry. After all, the self-assembly technique of the abalone has been successfully copied in the manufacture of a nanocomposite coating that is analogous to the nacre of its shell¹¹.

Andrew R. Parker, Victoria L. Welch, Dominique Driver, Natalia Martini

Department of Zoology, University of Oxford, Oxford OX1 3PS, UK

e-mail: andrew.parker@zoo.ox.ac.uk

1. Schultz, T. D. *Bull. Ent. Soc. Am.* **32**, 142–146 (1986).
2. Parker, A. R., McKenzie, D. R. & Large, M. C. *J. Exp. Biol.* **201**, 1307–1313 (1998).
3. Parker, A. R. *J. Opt. A* **2**, 15–28 (2000).
4. Sanders, J. V. *Nature* **204**, 1151–1153 (1964).
5. Ajgaonkar, M., Zhang, Y., Grebel, H. & Brown, R. A. *J. Opt. Soc. Am. B* **19**, 1391–1395 (2002).

6. Sanders, J. V. & Darragh, P. J. *Min. Rec.* **2**, 261–268 (1971).
7. Philipse, A. P. *J. Mat. Sci. Lett.* **8**, 1371–1373 (1989).
8. Land, M. F. *Progr. Biophys. Mol. Biol.* **24**, 75–106 (1972).
9. Parker, A. R., McPhedran, R. C., McKenzie, D. R., Botten, L. C. & Nicorovici, N.-A. *Nature* **409**, 36–37 (2001).
10. Nedelec, F. J., Surrey, T., Maggs, A. C. & Leibler, S. *Nature* **389**, 305–308 (1997).
11. Blackwell, J. & Weih, M. A. in *Chitin, Chitosan and Related Enzymes* (ed. Zikakis, J. P.) 257–272 (Academic, New York, 1984).

Competing financial interests: declared none.

COMMUNICATIONS ARISING

Plasma antioxidants

Health benefits of eating chocolate?

In assessing whether or not a compound or food acts as an antioxidant *in vivo*, a conventional approach is to monitor biological markers of oxidative damage in response to the intervention¹. Another is to measure the change in total plasma antioxidant capacity, as investigated by Serafini *et al.*² in relation to the consumption of chocolate in the presence and absence of milk. The implications of the authors' finding that eating chocolate causes an increase in total plasma antioxidant capacity, and the mechanism by which this is achieved, must also be considered — however, it should not be assumed that the effect is necessarily beneficial.

I calculate that the maximum plasma epicatechin concentration in the study of Serafini *et al.* is about 1 μ M. Epicatechin metabolites are likely to be present at lower concentrations and to have reduced antioxidant activity compared with that of epicatechin itself, because of blocking of radical-scavenging hydroxyl groups by conjugation³. However, the total plasma antioxidant capacity (TPAC) measured by Serafini *et al.*² rose by up to 18%. When measured by the ferric-reducing antioxidant-potential (FRAP) used by the authors, TPAC is usually 0.6–1.6 mM (ref. 4), of which 18% would be 108–288 μ M. The rise in TPAC is therefore so large that it is unlikely to be due in significant part to the antioxidant action of epicatechin and its metabolites, or of other phenolics in chocolate.

The FRAP activity of human plasma is mainly attributable to ascorbate, α -tocopherol, bilirubin and urate⁴. Given normal plasma levels of these substances⁵, an increase in urate concentration is most likely to account for the results of Serafini *et al.*, because urate is present in plasma at much higher levels than those of other antioxidants, and chocolate is unlikely to contain much ascorbate. Although α -tocopherol may be present in chocolate, even a high intake of this vitamin can increase its concentration in plasma by only a few micromolar at most⁵.

The mechanism and consequences of increased plasma urate levels following consumption of chocolate would be interesting to investigate, but should not necessarily be

regarded as beneficial. Hyperuricaemia has been associated with stroke, cardiovascular and renal morbidity, and gout^{5–8}.

Evidence is mounting regarding the potential importance of the overall redox network in disease prevention. It is crucial to our understanding to elucidate the mechanisms by which TPAC is modulated, and the effect of food on this important parameter.

Barry Halliwell

Department of Biochemistry, National University of Singapore, 8 Medical Drive, MD7 #03-08, 117597 Singapore

e-mail: bchbh@nus.edu.sg

1. Halliwell, B. *Nutr. Rev.* **57**, 104–111 (1999).
2. Serafini, M. *et al. Nature* **424**, 1013 (2003).
3. Rechner, A. R. *et al. Free Radic. Res.* **36**, 1229–1241 (2002).
4. Benzie, I. F. F. & Strain, J. J. *Anal. Biochem.* **239**, 70–76 (1996).
5. Halliwell, B. & Gutteridge, J. M. C. *Arch. Biochem. Biophys.* **280**, 1–8 (1990).
6. Weir, C. J. *et al. Stroke* **34**, 1951–1956 (2003).
7. Rott, K. T. *et al. J. Am. Med. Assoc.* **289**, 2857–2860 (2003).
8. Johnson, R. J. *et al. Hypertension* **41**, 1189–1190 (2003).

Nutrition

Milk and absorption of dietary flavanols

Flavanol compounds in wine, cocoa products and tea can exert a cardioprotective effect, for example by influencing endothelial-cell function¹, antithrombotic mechanisms² and blood pressure^{3,4}. Serafini *et al.*⁵ claim that consuming dark chocolate, but not milk chocolate or dark chocolate together with milk, increases the antioxidant capacity of human plasma, and suggest that interaction between milk proteins and chocolate flavonoids inhibits the *in vivo* antioxidant activity of chocolate and the absorption of epicatechin into the bloodstream. This inference could have implications beyond chocolate consumption if dairy products do indeed counteract the putative health benefits of dietary flavanols.

The results of Serafini *et al.*⁵ are open to a different interpretation if the biological availability and subsequent activity of any compound depends on the varying nutritional and biophysical properties of the matrix in which it is ingested (that is, caloric background, lipid/water content, viscosity, density, extent of mastication). To compare the absorption of epicatechin for chocolate ingested in the presence and absence of milk, it is necessary to control for the composition of the matrix in which the flavanols are delivered.

Compared with values obtained after consumption of dark chocolate alone (100 g, of which about 30 g is lipid), Serafini *et al.* find a lower plasma antioxidant capacity, as well as a reduction in the area under the curve of a plot of plasma antioxidant activity for epicatechin against time, in the case of milk chocolate (200 g, of which about 60 g is

lipid) and for dark chocolate consumed with milk (100 g plus 200 ml milk, corresponding to more than 30 g lipid). These differences may therefore reflect matrix-dependent effects on flavanol absorption.

A mechanistic link between the antioxidant properties of flavonoids *in vitro* and their biological activity *in vivo* is not properly established, partly because factors such as biological transformation and tissue/plasma concentrations are often not considered⁶. Measurements of plasma antioxidant capacity ideally need to be complemented with markers of cardiovascular function to assess the biological effect of flavanols on cardiovascular health *in vivo*.

We tested the possible inhibition by milk of the absorption of chocolate flavanols in 12 healthy, non-smoking volunteers whose blood-lipid profiles were normal and who were not taking any dietary supplement (the study was approved by the Internal Review Board of the University of California, Davis, and informed consent was obtained from all subjects). Participants consumed a cocoa beverage (containing 0.66 g cocoa solids, the equivalent of 5.44 mg monomeric flavanols and 15.65 mg proanthocyanidins, per kg body weight) prepared with either whole milk (3.25% lipid) or water (supplemented with carbohydrate and 3.25% lipid as a control). Both beverages delivered an equal amount of fluid (4.8 g per kg body weight).

The area under the curve (AUC) in plots of epicatechin plasma concentration against time (where plasma epicatechin was extracted⁷ and analysed by high-performance liquid chromatography⁸) shows that there is no significant difference (analysis of variance (ANOVA), $P = 0.499$) in epicatechin concentration in plasma after the consumption of a milk-containing (mean AUC, $3,074.8 \pm 536.7$ ng ml⁻¹ h⁻¹) or water-based (mean AUC, $2,580.5 \pm 655.4$ ng ml⁻¹ h⁻¹) cocoa beverage under isocaloric and isolipidaemic conditions.

Bearing in mind that measurements of plasma antioxidant capacity must be interpreted with caution in evaluating flavonoid-mediated biological effects, we used the total antioxidant potential (TRAP) assay⁹ as an independent test of the results of Serafini *et al.*⁵, who measured the ferric-reducing antioxidant potential (FRAP). We found that consumption of a cocoa-containing beverage resulted in a statistically significant, milk-independent increase in plasma antioxidant capacity, as shown by comparing the baseline value (mean, 234.2 ± 12.47 nmol trolox (vitamin E equivalent) per ml) with the maximum values for the milk-containing (mean maximum, 291.4 ± 8.6 nmol trolox ml⁻¹; ANOVA, $P < 0.001$) and water-containing (mean maximum, 283.72 ± 21.6 nmol trolox ml⁻¹; ANOVA, $P < 0.001$) drinks.

Results based on AUC values for plasma antioxidant capacity plotted against time, which take into account temporal differences in absorption, also indicate that milk does

not influence the cocoa-mediated increase in plasma antioxidant capacity (mean AUC_{milk}, $1,028.7 \pm 40.5$; mean AUC_{water}, $1,005.6 \pm 35.7$ nmol trolox equivalents ml⁻¹ h⁻¹; t -test, $P \pm 0.369$). Furthermore, the consumption of a milk-containing cocoa beverage resulted in a 25–30% reduction in platelet-mediated haemostasis (R.R.H. and C.L.K., unpublished results), a functional marker that is related to cardiovascular health.

Our findings show that the presence of milk in cocoa products does not counteract the absorption and biological activity of monomeric flavanols from cocoa products, nor does it affect plasma antioxidant capacity. In the context of potential health benefits mediated by dietary flavanols, food-matrix-dependent, temporal effects on absorption must be taken into account when comparing different food formulations and processing methods in terms of the biological activity of their constituent compounds.

Hagen Schroeter*, Roberta R. Holt*, Timothy J. Orozco*, Harold H. Schmitz†, Carl L. Keen*‡

Departments of *Nutrition and ‡Internal Medicine, University of California, Davis, California 95616, USA

e-mail: clkeen@ucdavis.edu

†Analytical and Applied Sciences, Masterfoods USA, Hackettstown, New Jersey 07840, USA

1. Heiss, C. *et al.* *J. Am. Med. Assoc.* **290**, 1030–1031 (2003).
2. Holt, R. R., Schramm, D. D., Keen, C. L., Lazarus, S. A. & Schmitz, H. H. *J. Am. Med. Assoc.* **287**, 2212–2213 (2002).
3. Corder, R. *et al.* *Nature* **414**, 863–864 (2001).
4. Taubert, D., Berkels, R., Roosen, R. & Klaus, W. *J. Am. Med. Assoc.* **290**, 1029–1030 (2003).
5. Serafini, M. *et al.* *Nature* **424**, 1013 (2003).
6. Spencer, J. P., Schroeter, H., Rechner, A. R. & Rice-Evans, C. *Antioxid. Redox Signal* **3**, 1023–1039 (2001).
7. Abd El Mohsen, M. M. *et al.* *Free Rad. Biol. Med.* **33**, 1693–1702 (2002).
8. Holt, R. R. *et al.* *Am. J. Clin. Nutr.* **76**, 798–804 (2002).
9. Lissi, E., Salim-Hanna, M., Pascual, C. & del Castillo, M. D. *Free Rad. Biol. Med.* **18**, 153–158 (1995).

Competing financial interests: declared; see online version for details.

Serafini *et al.* reply — Our results indicate that there is an increase in total antioxidant capacity (TAC) and (–)epicatechin content of plasma in people who have consumed dark chocolate, and that these effects are reduced by the presence of milk¹. Halliwell and Schroeter *et al.* raise issues that are central to the debate over the fate and potential protective effects of dietary antioxidants.

Schroeter *et al.* suggest that a matrix effect in our study may have delayed the absorption of antioxidants when chocolate was consumed with milk, and that this went unnoticed because our experiments were terminated after 4 h. However, this is unlikely for several reasons. All of the participants in our trials fasted overnight for 12 h and, after supplementation with chocolate/milk, no other product was consumed for the duration of the experiment. Individuals who ate dark chocolate showed a significant increase in plasma TAC

after 1 h and in plasma (–)epicatechin for 4 h after ingestion (as measured by the area under the curve in Fig. 1a of ref. 1). In a pilot study, we also monitored plasma TAC for 6 h (results not shown) and found no increase following the consumption of either milk chocolate or dark chocolate with milk.

The study described by Schroeter *et al.* differs in key respects from ours. The authors used a chocolate beverage, rather than solid chocolate. The amount of milk consumed with the beverage is not stated but, on the basis of the lipid content (3% as opposed to 30%), it is substantially less than in our study, which could explain why they did not observe an inhibition of plasma TAC by milk. Another factor could be that Schroeter *et al.* measured plasma TAC by using the TRAP-luminol assay, which is unlike the FRAP assay that we used in that it measures radical-scavenging activity rather than reducing power, and as such determines a different feature of TAC².

We were careful to distinguish between TAC and (–)epicatechin absorption because the aim of our investigation was to evaluate the effect of food associations on plasma TAC, and not to identify the compounds responsible for the effect. We agree that the increase in plasma TAC cannot be explained solely on the basis of increased (–)epicatechin levels. Other chocolate polyphenols, such as procyanidins or their *in vivo* metabolites, could also be involved in increasing FRAP values. However, whatever the components responsible, they too are likely to be affected by milk, causing a reduction *in vivo* in the antioxidant properties of dark chocolate.

We measured urate concentrations in plasma under similar conditions (results not shown) and the results are commensurate with an involvement of urate in the redox network. However, as noted by Halliwell, the situation is not straightforward and we did not therefore speculate on the contribution of urate in this context.

TAC measurement is increasingly being used to monitor redox status *in vivo* as well as for epidemiological purposes³. Understanding the factors that regulate plasma TAC will help to clarify the interplay between plant-derived foods, plasma redox status and oxidative disease; the effect of food associations on TAC *in vivo* is just one, albeit complicated, variable.

Mauro Serafini*, Alan Crozier†

*Antioxidant Research Laboratory, Unit of Human Nutrition, National Institute for Food and Nutrition Research, 00178 Rome, Italy
e-mail: serafini@inran.it

†Plant Products and Human Nutrition Group, Division of Biochemistry and Molecular Biology, Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK

1. Serafini, M. *et al.* *Nature* **424**, 1013 (2003).
2. Del Rio, D., Serafini, M. & Pellegrini, N. *Nutr. Met. Card. Dis.* **12**, 343–351 (2002).
3. Serafini, M., Bellocco, R., Wolk, A. & Ekström, A. M. *Gastroenterology* **123**, 985–991 (2002).

The International HapMap Project

The International HapMap Consortium*

*Lists of participants and affiliations appear at the end of the paper

The goal of the International HapMap Project is to determine the common patterns of DNA sequence variation in the human genome and to make this information freely available in the public domain. An international consortium is developing a map of these patterns across the genome by determining the genotypes of one million or more sequence variants, their frequencies and the degree of association between them, in DNA samples from populations with ancestry from parts of Africa, Asia and Europe. The HapMap will allow the discovery of sequence variants that affect common disease, will facilitate development of diagnostic tools, and will enhance our ability to choose targets for therapeutic intervention.

Common diseases such as cardiovascular disease, cancer, obesity, diabetes, psychiatric illnesses and inflammatory diseases are caused by combinations of multiple genetic and environmental factors¹. Discovering these genetic factors will provide fundamental new insights into the pathogenesis, diagnosis and treatment of human disease. Searches for causative variants in chromosome regions identified by linkage analysis have been highly successful for many rare single-gene disorders. By contrast, linkage studies have been much less successful in locating genetic variants that affect common complex diseases, as each variant individually contributes only modestly to disease risk^{2,3}. A complementary approach to identifying these specific genetic risk factors is to search for an association between a specific variant and a disease, by comparing a group of affected individuals with a group of unaffected controls⁴. In the absence of strong natural selection, there is likely to be a broad spectrum of frequency of such variants, many of which are likely to be common in the population. A number of association studies, focused on candidate genes, regions of linkage to a disease or more large-scale surveys, have already led to the discovery of genetic risk factors for common diseases. Examples include type 1 diabetes (human leukocyte antigen (HLA)⁵), insulin⁶ and *CTLA4* (ref. 7)), Alzheimer's disease (*APOE*)⁸, deep vein thrombosis (factor V)⁹, inflammatory bowel disease (*NOD2* (refs 10, 11) and also 5q31 (ref. 12)), hypertriglyceridaemia (*APOA5*)¹³, type 2 diabetes (*PPARG*)^{14,15}, schizophrenia (neuregulin 1)¹⁶, asthma (*ADAM33*)¹⁷, stroke (*PDE4D*)¹⁸ and myocardial infarction (*LTA*)¹⁹.

One approach to doing association studies involves testing each putative causal variant for correlation with the disease (the 'direct' approach)². To search the entire genome for disease associations would entail the substantial expense of whole-genome sequencing of numerous patient samples to identify the candidate variants³. At present, this approach is limited to sequencing the functional parts of candidate genes (selected on the basis of a previous functional or genetic hypothesis) for potential disease-associated candidate variants. An alternative approach (the 'indirect' approach) has been proposed²⁰, whereby a set of sequence variants in the genome could serve as genetic markers to detect association between a particular genomic region and the disease, whether or not the markers themselves had functional effects. The search for the causative variants could then be limited to the regions showing association with the disease.

Two insights from human population genetics suggest that the indirect approach is able to capture most human sequence variation, with greater efficiency than the direct approach. First, ~90% of sequence variation among individuals is due to common variants²¹. Second, most of these originally arose from single historical mutation events, and are therefore associated with nearby variants that were present on the ancestral chromosome on which the mutation occurred. These associations make the indirect approach

feasible to study variants in candidate genes, chromosome regions or across the whole genome. Prior knowledge of putative functional variants is not required. Instead, the approach uses information from a relatively small set of variants that capture most of the common patterns of variation in the genome, so that any region or gene can be tested for association with a particular disease, with a high likelihood that such an association will be detectable if it exists.

The aim of the International HapMap Project is to determine the common patterns of DNA sequence variation in the human genome, by characterizing sequence variants, their frequencies, and correlations between them, in DNA samples from populations with ancestry from parts of Africa, Asia and Europe. The project will thus provide tools that will allow the indirect association approach to be applied readily to any functional candidate gene in the genome, to any region suggested by family-based linkage analysis, or ultimately to the whole genome for scans for disease risk factors.

Common variants responsible for disease risk will be most readily approached by this strategy, but not all predisposing variants are common. However, it should be noted that even a relatively uncommon disease-associated variant can potentially be discovered using this approach. Reflecting its historical origins, the uncommon variant will be travelling on a chromosome that carries a characteristic pattern of nearby sequence variants. In a group of people affected by a disease, the rare variant will be enriched in frequency compared with its frequency in a group of unaffected controls. This observation, for example, was of considerable assistance in the identification of the genes responsible for cystic fibrosis²² and diastrophic dysplasia²³, after linkage had pointed to the general chromosomal region.

Below we provide a brief description of human sequence variation, and then describe the strategy and key components of the project. These include the choice of samples and populations for study, the process of community engagement or public consultation, selection of single-nucleotide polymorphisms (SNPs), genotyping, data release and analysis.

Human DNA sequence variation

Any two copies of the human genome differ from one another by approximately 0.1% of nucleotide sites (that is, one variant per 1,000 bases on average)^{24–27}. The most common type of variant, a SNP, is a difference between chromosomes in the base present at a particular site in the DNA sequence (Fig. 1a). For example, some chromosomes in a population may have a C at that site (the 'C allele'), whereas others have a T (the 'T allele'). It has been estimated that, in the world's human population, about 10 million sites (that is, one variant per 300 bases on average) vary such that both alleles are observed at a frequency of $\geq 1\%$, and that these 10 million common SNPs constitute 90% of the variation in the popu-

lation^{21,28}. The remaining 10% is due to a vast array of variants that are each rare in the population. The presence of particular SNP alleles in an individual is determined by testing ('genotyping') a genomic DNA sample.

Nearly every variable site results from a single historical mutational event as the mutation rate is very low (of the order of 10^{-8} per site per generation) relative to the number of generations since the most recent common ancestor of any two humans (of the order of 10^4 generations). For this reason, each new allele is initially associated with the other alleles that happened to be present on the particular chromosomal background on which it arose. The specific set of alleles observed on a single chromosome, or part of a chromosome, is called a haplotype (Fig. 1b). New haplotypes are formed by additional mutations, or by recombination when the maternal and paternal chromosomes exchange corresponding segments of DNA, resulting in a chromosome that is a mosaic of the two parental haplotypes²⁹.

The coinheritance of SNP alleles on these haplotypes leads to associations between these alleles in the population (known as linkage disequilibrium, LD). Because the likelihood of recombination between two SNPs increases with the distance between them, on average such associations between SNPs decline with distance. Many empirical studies have shown highly significant levels of LD, and often strong associations between nearby SNPs, in the human genome^{30–34}. These strong associations mean that in many chromosome regions there are only a few haplotypes, and these account for most of the variation among people in those regions^{31,35,36}.

The strong associations between SNPs in a region have a practical value: genotyping only a few, carefully chosen SNPs in the region will provide enough information to predict much of the information about the remainder of the common SNPs in that region. As a result, only a few of these 'tag' SNPs are required to identify each of

the common haplotypes in a region^{35,37–39} (Fig. 1c).

As the extent of association between nearby markers varies dramatically across the genome^{30–32,34,35,40}, it is not efficient to use SNPs selected at random or evenly spaced in the genome sequence. Instead, the patterns of association must be empirically determined for efficient selection of tag SNPs. On the basis of empirical studies, it has been estimated that most of the information about genetic variation represented by the 10 million common SNPs in the population could be provided by genotyping 200,000 to 1,000,000 tag SNPs across the genome^{31,36,38,39}. Thus, a substantial reduction in the amount of genotyping can be obtained with little loss of information, by using knowledge of the LD present in the genome.

For common SNPs, which tend to be older than rare SNPs, the patterns of LD largely reflect historical recombination and demographic events⁴¹. Some recombination events occur repeatedly at 'hotspots'^{30,42}. The result of these processes is that current chromosomes are mosaics of ancestral chromosome regions²⁹. This explains the observations that haplotypes and patterns of LD are shared by apparently unrelated chromosomes within a population and generally among populations⁴³.

These observations are the conceptual and empirical foundation for developing a haplotype map of the human genome, the 'HapMap'. This map will describe the common patterns of variation, including associations between SNPs, and will include the tag SNPs selected to most efficiently and comprehensively capture this information.

The International HapMap Consortium

An initial meeting to discuss the scientific and ethical issues associated with developing a human haplotype map was held in Washington DC on 18–19 July 2001 (<http://www.genome.gov/10001665>). Groups were organized to consider the ethical issues, to develop the scientific plan and to choose the populations to

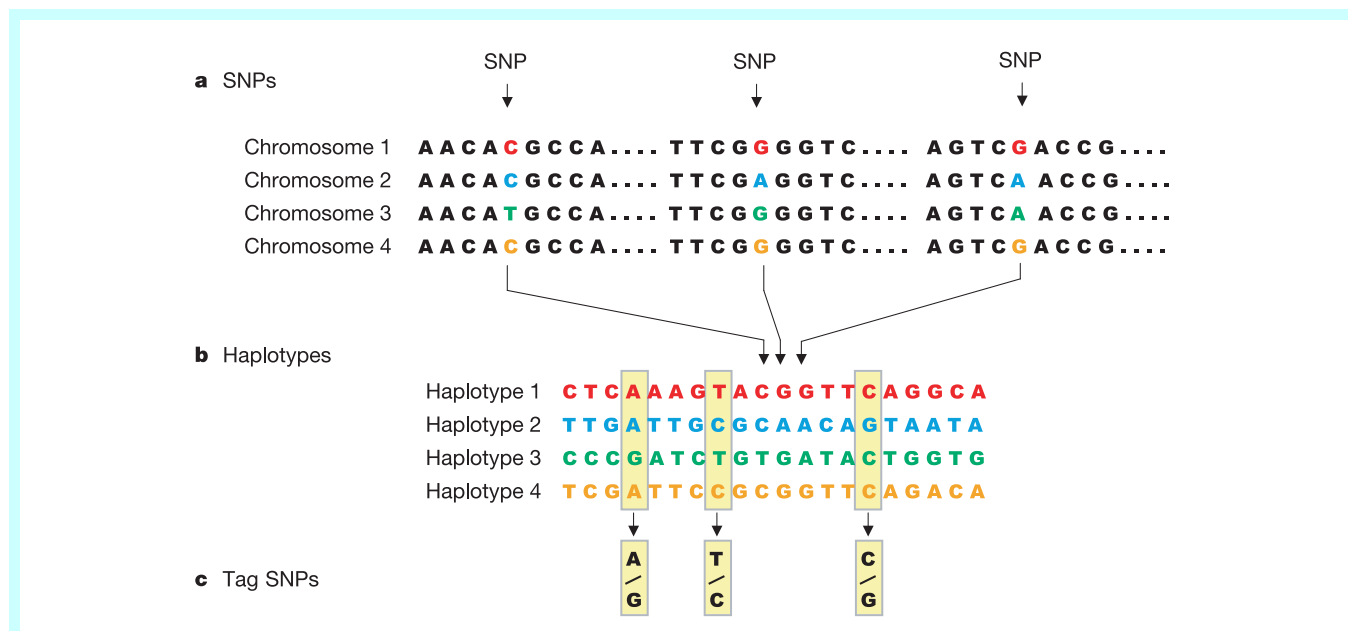


Figure 1 SNPs, haplotypes and tag SNPs. **a**, SNPs. Shown is a short stretch of DNA from four versions of the same chromosome region in different people. Most of the DNA sequence is identical in these chromosomes, but three bases are shown where variation occurs. Each SNP has two possible alleles; the first SNP in panel **a** has the alleles C and T. **b**, Haplotypes. A haplotype is made up of a particular combination of alleles at nearby SNPs. Shown here are the observed genotypes for 20 SNPs that extend across 6,000 bases of DNA. Only the variable bases, including the

three SNPs that are shown in panel **a**. For this region, most of the chromosomes in a population survey turn out to have haplotypes 1–4. **c**, Tag SNPs. Genotyping just the three tag SNPs out of the 20 SNPs is sufficient to identify these four haplotypes uniquely. For instance, if a particular chromosome has the pattern A–T–C at these three tag SNPs, this pattern matches the pattern determined for haplotype 1. Note that many chromosomes carry the common haplotypes in the population.

include. The International HapMap Project (<http://www.hapmap.org/>) was then formally initiated with a meeting in Washington DC on 27–29 October 2002 (<http://www.genome.gov/10005336>). The participating groups and funding sources are listed in Table 1.

DNA samples and populations

Human populations are the products of numerous social, historical and demographic processes. As a result, no populations are typical, special or sharply bounded^{44,45}. As most common patterns of variation can be found in any population⁴⁶, no one population is essential for inclusion in the HapMap. Nonetheless, we decided to include several populations from different ancestral geographic locations to ensure that the HapMap would include most of the common variation and some of the less common variation in different populations, and to allow examination of various hypotheses about patterns of LD.

Studies of allele frequency distributions suggest that ancestral geography is a reasonable basis for sampling human populations^{44,47,48}. Pilot studies using samples from the Yoruba, Japanese, Chinese and individuals with ancestry from Northern and Western Europe have shown substantial similarity in their haplotype patterns, although the frequencies of haplotypes often differ^{31,44}. Given these scientific findings, coupled with consideration of ethical, social and cultural issues, these populations were approached for inclusion in the HapMap through a process of community engagement or consultation (see Box 1).

The HapMap developed with samples from these four large

populations will include a substantial amount of the genetic variation found in all populations throughout the world. The goal of the HapMap is medical, and the common patterns of variation identified by the project will be useful to identify genes that contribute to disease and drug response in many other populations. Samples from several other populations are being collected for studies that will examine how similar their haplotype patterns are to those in the HapMap. If the patterns found are very different, samples from some of these populations may be genotyped on a large scale to make the HapMap more applicable to them. Further follow-up studies in other populations, small and large, are likely to be undertaken by scientists in many nations for common disease gene discovery.

The project will study a total of 270 DNA samples: 90 samples (see Supplementary Information, part 1) from a US Utah population with Northern and Western European ancestry (samples collected in 1980 by the Centre d'Etude du Polymorphisme Humain (CEPH)⁴⁹ and used for other human genetic maps, 30 trios of two parents and an adult child), and new samples collected from 90 Yoruba people in Ibadan, Nigeria (30 trios), 45 unrelated Japanese in Tokyo, Japan, and 45 unrelated Han Chinese in Beijing, China. All donors gave specific consent for their inclusion in the project. Population membership was determined in ways appropriate for each culture: for the Yoruba by asking the donor whether all four grandparents were Yoruba, for the Han Chinese by asking the donor whether at least three of four grandparents were Han Chinese, and for the Japanese by self-identification. The CEPH samples are available

Table 1 Groups participating in the International HapMap Project

Country	Research group		Institution	Role	Per cent genome		Chromosomes	Funding agency
Japan	Yusuke Nakamura		RIKEN, Univ. of Tokyo	Genotyping: Third Wave	25.1%		5, 11, 14, 15, 16, 17, 19	Japanese MEXT
	Ichiro Matsuda		Health Sciences Univ. of Hokkaido, Eubios Ethics Inst., Shinshu Univ.	Public consultation, Samples				
United Kingdom	David Bentley		Sanger Inst.	Genotyping: Illumina	24.0%		1, 6, 10, 13, 20	Wellcome Trust
	Peter Donnelly		Univ. of Oxford	Analysis				TSC, US NIH
	Lon Cardon		Univ. of Oxford	Analysis				Wellcome Trust, TSC, US NIH
Canada	Thomas Hudson		McGill Univ. and G��n��me Qu��bec Innovation Centre	Genotyping: Illumina	10.0%		2, 4p	Genome Canada, G��n��me Qu��bec
China	Huaming Yang The Chinese HapMap Consortium	Changqing Zeng	Beijing Genomics Inst., Chinese National Human Genome Center at Beijing	Genotyping: PerkinElmer and Sequenom	4.8%	10%	3, 8p, 21	Chinese MOST, Chinese Academy of Sciences, Natural Science Foundation of China, Hong Kong Innovation and Technology Commission, University Grants Committee of Hong Kong
		Wei Huang	Chinese National Human Genome Center at Shanghai, Inst. of Biomedical Sciences (Taiwan)	Genotyping: Illumina	3.2%			
		Lap-Chee Tsui	Univ. of Hong Kong, Hong Kong Univ. of Sci. & Tech., Chinese Univ. of Hong Kong	Genotyping: Sequenom	2.0%			
	Houcun Zhang		Beijing Normal Univ.	Community engagement				Chinese MOST
	Changqing Zeng		Beijing Genomics Inst.	Samples				
	United States	Mark Chee		Illumina	Genotyping: Illumina	15.5%	30.9%	8q, 9, 18q, 22, X
David Altshuler		Whitehead Inst.	Genotyping: Sequenom and Illumina	9.1%	4q, 7q, 18p, Y			
			Analysis					
Richard Gibbs		Baylor College of Medicine, ParAllele	Genotyping: ParAllele	4.4%	12			
Pui-Yan Kwok		UCSF, Washington Univ.	Genotyping: PerkinElmer	1.9%	7p			
Aravinda Chakravarti		Johns Hopkins Univ.	Analysis					
Mark Leppert		Univ. of Utah	Community engagement Samples				W.M. Keck Foundation, Delores Dore Eccles Foundation, US NIH	
Nigeria	Charles Rotimi		Howard Univ., Univ. of Ibadan	Community engagement Samples				US NIH
	Lincoln Stein		Cold Spring Harbor Lab., New York	Data coordination centre				TSC

from the non-profit Coriell Institute of Medical Research (<http://locus.umdnj.edu/nigms/>); cell lines and DNA from the new samples will be available from Coriell in early 2004 for future studies with research protocols approved by appropriate ethics committees. It is anticipated that other researchers will genotype additional SNPs in these samples in the future, and that these data will continuously improve the HapMap.

These samples will have population and sex identifiers without information that could link them to individual donors. As the goal of the project is solely to identify patterns of genetic variation, no medical or other phenotypic information will be included. About 50% more samples were collected than will be used, so that inclusion of a sample from any particular donor cannot be known.

Samples of 45 unrelated individuals should be sufficient to find 99% of haplotypes with a frequency of 5% or greater in a population. Studies of LD can use random individual samples, trios or larger pedigrees; each design has advantages (ease of sampling) and disadvantages (decreasing efficiency with increasing numbers of related individuals). Analysis of existing data and computer simulations suggested that unrelated individuals and trios have considerable power for estimating local LD patterns. The trios will provide useful information on the accuracy of the genotyping platforms being used for the project.

Choice of SNPs

A high density of SNPs is needed to describe adequately the genetic variation across the entire genome. When the project started, the average density of markers in the public database dbSNP (<http://www.ncbi.nlm.nih.gov/SNP/>)⁵⁰ was approximately one every kilobase (2.8 million SNPs) but, given their variable distribution, many regions had a lower density of SNPs.

Further SNPs were obtained by random shotgun sequencing from whole-genome and whole-chromosome (flow-sorted) libraries⁵¹, using methods developed for the initial human SNP map⁵², and also by collaboration with Perlegen Sciences³⁶ and through the purchase of sequence traces from Applied Biosystems⁵³ for SNP detection (see Supplementary Information, part 2). One useful result of this search for more SNPs is the confirmation of SNPs found previously. SNPs for which each allele has been seen independently in two or more samples ('double-hit' SNPs) have a higher average minor allele frequency than do 'single-hit' SNPs²⁸. This leads to substantial savings in assay development. On 4 November 2003, the number of SNPs (with a unique genomic position) in dbSNP (build 118) was 5.7 million, and the number of double-hit SNPs was over 2 million. By February of 2004, 6.8 million SNPs (with a unique genomic position) are expected to be in dbSNP and available for the project, including 2.7 million double-hit SNPs.

As the extent of LD and haplotypes varies by 100-fold across the genome^{30–32,34,35}, a hierarchical genotyping strategy has been adopted. In an initial round of genotyping, the project aims to genotype successfully 600,000 SNPs spaced at approximately 5-kilobase intervals and each with a minor allele frequency of at least 5%, in the 270 DNA samples. Priority is being given to previously validated SNPs, double-hit SNPs and SNPs causing amino-acid changes (as these may alter protein function). When these genotyping data are produced (by mid-2004; see below for details of data release), they will be analysed for associations between neighbouring SNPs. Additional SNPs will then be genotyped in the same DNA samples at a higher density only in regions where the associations are weak. Further rounds of analysis and genotyping will be carried out as required. It is expected that more than one million SNPs will be genotyped overall. This hierarchical strategy will permit regions of the genome with the least LD to be characterized at densities of up to one SNP per kilobase, maximizing the characterization of regions with associations only over short distances.

Genotyping

Each genotyping centre is responsible for genotyping all the samples for all the selected SNPs on the chromosome regions allocated (Table 1). Among the centres, a total of five high-throughput genotyping technologies are being used, which will provide an opportunity to compare their accuracy, success rate, throughput and cost. Access to several platforms is an advantage for the project, as a SNP assay that fails on one platform may be developed successfully using another method in order to fill a gap in the HapMap. All platforms will be evaluated using a common set of performance criteria to ensure that the quality of data produced for the project meets a uniformly high standard.

Genotype quality is being assessed in three ways. First, at the beginning of the project, all centres were assigned the same randomly selected set of 1,500 SNPs for assay development and genotyping in the 90 CEPH DNA samples being used for the project. Genotyping centres produced data that were on average more than 99.2% complete and more than 99.5% accurate (as compared to the consensus of at least two other platforms). Second, every genotyping experiment includes samples for internal quality checks, with each 96-well plate containing duplicates of five different samples, and one blank. In addition, the data from trios provide a check for consistent mendelian inheritance of SNP alleles. For all the populations, the data from the unrelated samples provide a check that the SNPs are in Hardy–Weinberg equilibrium (a test of genetic mating

Box 1

Community engagement, public consultation and individual consent

As no personally identifiable information will be linked to the samples, the risk that an individual will be harmed by a breach of privacy, or by discrimination based on studies that use the HapMap, is minimal. However, because tag SNPs for future disease studies will be chosen on the basis of haplotype frequencies in the populations included in the HapMap, the data will be identified as coming from one of the four populations involved, and it will be possible to make comparisons between the populations. As a result, the use of population identifiers may create risks of discrimination or stigmatization, as might occur if a higher frequency of a disease-associated variant were to be found in a group and this information were then overgeneralized to all or most of its members⁶⁴. It is possible that there are other culturally specific risks that may not be evident to outsiders⁶⁵. To identify and address these group risks, a process of community engagement, or public consultation, was undertaken to confer with members of the populations being approached for sample donation about the implications of their participation in the project^{66,67}. The goal was to give people in the localities where donors were recruited the opportunity to have input into the informed consent and sample collection processes, and into such issues as how the populations from which the samples were collected would be named. Community engagement is not a perfect process, but it is an effort to involve potential donors in a more extended consideration of the implications of a research project before being asked to take part in it⁶⁸. Community engagement and individual informed consent were conducted under the auspices of local governments and ethics committees, taking into account local ethical standards and international ethical guidelines. As in any cross-cultural endeavour, the form and outcome of the processes varied from one population to another. A Community Advisory Group is being set up for each community to serve as a continuing liaison with the sample repository, to ensure that future uses of the samples are consistent with the uses described in the informed consent documents. A more detailed article discussing ethical, social and cultural issues relevant to the project, and describing the processes used to engage donor populations in identifying and evaluating these issues, is in preparation.

patterns). Although a small proportion of SNPs may fail these checks for biological reasons, they more typically fail if a genotyping platform makes consistent errors, such as undercalling heterozygotes. Third, a sample of SNP genotypes deposited by each centre will be selected at random and re-genotyped by other centres. These stringent third-party evaluations of quality will ensure the completeness and reliability of the data produced by the project.

Data release

The project is committed to rapid and complete data release, and to ensuring that project data remain freely available in the public domain at no cost to users. The project follows the data-release principles of a 'community resource project' (<http://www.wellcome.ac.uk/en/1/awtpubrepdat.html>).

All data on new SNPs, assay conditions, and allele and genotype frequencies will be released rapidly into the public domain on the internet at the HapMap Data Coordination Center (DCC) (<http://www.hapmap.org/>) and deposited in dbSNP. Individual genotype and haplotype data initially will be made available at the DCC under a short-term 'click-wrap' licence agreement. This strategy has been adopted to ensure that data from the project cannot be incorporated into any restrictive patents, and will thus remain freely available in the long term. The only condition for data access is that users must agree not to restrict use of the data by others and to share the data only with others who have agreed to the same condition. When haplotypes are defined in a region, then the individual genotypes, haplotypes and tag SNPs in that region will be publicly released to dbSNP, where there are no licensing conditions. Project participants have agreed that their own laboratories will access the data through the DCC and under the click-wrap licence, ensuring that all scientists have equal access to the data for research.

The consortium believes that SNP, genotype and haplotype data in the absence of specific utility do not constitute appropriately patentable inventions. Specific utility would involve, for example, finding an association of a SNP or haplotype with a medically important phenotype such as a disease risk or drug response. The project does not include any phenotype association studies. However, the data-release policy does not block users from filing for appropriate intellectual property on such associations, as long as any ensuing patent is not used to prevent others' access to the HapMap data.

Data analysis

The project will apply existing and new methods for analysis and display of the data. LD between pairs of markers will be calculated using standard measures such as D' (ref. 54), r^2 (refs 55, 56) and others. Various methods are being evaluated to define regions of high LD and haplotypes along chromosomes. Existing methods include 'sliding window' LD profiles^{57,58}, LD unit maps⁵⁹, haplotype blocks^{31,35} and estimates of meiotic recombination rates along chromosomes^{35,60–62}. After analysis of the LD in the first phase of the project, regions in which there is little or no LD will be identified and ranked for further SNP selection and genotyping. Methods to select optimal collections of tag SNPs will be developed and evaluated (see above). The project will thus provide views of the data and tag SNPs that will be useful to the research community. As all data and analysis methods will be made available, other researchers will also be able to analyse the data and improve the analysis methods.

To assist optimization of SNP selection and analysis of LD and haplotypes, a pilot study is underway to produce a dense set of genotypes across large genomic regions. Ten 500-kilobase regions of the genome (see Supplementary Information, part 3) will be sequenced in 48 unrelated HapMap DNA samples (16 CEPH (currently being sequenced), 16 Yoruba, 8 Japanese and 8 Han Chinese). All SNPs identified, as well as any additional SNPs in the public databases, will be genotyped in all of the 270 HapMap DNA

samples, and the genotype data will be released following the guidelines described above. This study will provide dense genotype data for developing methods for SNP selection and for assessing the completeness of the information extracted, to guide the later stages of genotyping.

When the HapMap is used to examine large genomic regions, the problem of multiple comparisons will arise from testing tens to hundreds of thousands of SNPs and haplotypes for disease associations. This will lead to difficulty in separating true from false-positive results. Thus, new statistical methods, replication studies and functional analyses of variants will be important to confirm the findings and identify the functionally important SNPs.

Conclusion

The goal of the International HapMap Project is to develop a research tool that will help investigators across the globe to discover the genetic factors that contribute to susceptibility to disease, to protection against illness and to drug response. The HapMap will provide an important shortcut to carry out candidate-gene, linkage-based and genome-wide association studies, transforming an unfeasible strategy into a practical one. In its scope and potential consequences, the International HapMap Project has much in common with the Human Genome Project, which sequenced the human genome⁶³. Both projects have been scientifically ambitious and technologically demanding, have involved intense international collaboration, have been dedicated to the rapid release of data into the public domain, and promise to have profound implications for our understanding of human biology and human health. Whereas the sequencing project covered the entire genome, including the 99.9% of the genome where we are all the same, the HapMap will characterize the common patterns within the 0.1% where we differ from each other.

For the full potential of the HapMap to be realized, several things must occur. The technology for genotyping must become more cost efficient, and the analysis methods must be improved. Pilot studies with other populations must be completed to confirm that the HapMap is generally applicable, with consideration given to expanding the HapMap if needed so that all major world populations can derive the greatest benefit. To use the tools created by the HapMap, later projects must establish carefully phenotyped sets of affected and unaffected individuals for many common diseases in a way that preserves confidentiality but retains detailed clinical and environmental exposure data. Longitudinal cohort studies of hundreds of thousands of individuals will also be invaluable for assessing the genetic and environmental contributions to disease.

Careful and sustained attention must also be paid to the ethical issues that will be raised by the HapMap and the studies that will use it. By consulting members of donor populations about the consent process and the implications of population-specific findings before sample collection, the project has helped to advance the ethical standard for international population genetics research. Future population genetics projects will continue to refine this approach. It will be an ongoing challenge to avoid misinterpretations or misuses of results from studies that use the HapMap. Researchers using the HapMap should present their findings in ways that avoid stigmatizing groups, conveying an impression of genetic determinism, or attaching incorrect levels of biological significance to largely social constructs such as race.

The HapMap holds much promise as a powerful new tool for discovery—to enhance our understanding of the hereditary factors involved in health and disease. Realizing its full benefits will involve the close partnership of basic science researchers, population geneticists, epidemiologists, clinicians, social scientists, ethicists and the public. □

doi:10.1038/nature02168.

1. King, R. A., Rotter, J. I. & Motulsky, A. G. *The Genetic Basis of Common Diseases* Vol. 20 (eds Motulsky, A. G., Harper, P. S., Scriver, C. & Bobrow, M.) (Oxford Univ. Press, Oxford, 1992).
2. Risch, N. J. Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856 (2000).
3. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nature Genet.* **33** (Suppl.), 228–237 (2003).
4. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
5. Dorman, J. S., LaPorte, R. E., Stone, R. A. & Trucco, M. Worldwide differences in the incidence of type 1 diabetes are associated with amino acid variation at position 57 of the HLA-DQ β chain. *Proc. Natl Acad. Sci. USA* **87**, 7370–7374 (1990).
6. Bell, G. I., Horita, S. & Karam, J. H. A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes* **33**, 176–183 (1984).
7. Nisticò, L. *et al.* The *CTLA-4* gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes. *Hum. Mol. Genet.* **5**, 1075–1080 (1996).
8. Strittmatter, W. J. & Roses, A. D. Apolipoprotein E and Alzheimer's disease. *Annu. Rev. Neurosci.* **19**, 53–77 (1996).
9. Dahlbäck, B. Resistance to activated protein C caused by the factor V R³⁰⁶Q mutation is a common risk factor for venous thrombosis. *Thromb. Haemost.* **78**, 483–488 (1997).
10. Hugot, J. P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).
11. Ogura, Y. *et al.* A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* **411**, 603–606 (2001).
12. Rioux, J. D. *et al.* Genetic variation in the 5q31 cytokine gene cluster confers susceptibility to Crohn disease. *Nature Genet.* **29**, 223–228 (2001).
13. Pennacchio, L. A. *et al.* An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science* **294**, 169–173 (2001).
14. Deeb, S. S. *et al.* A Pro12Ala substitution in PPAR γ 2 associated with decreased receptor activity, lower body mass index and improved insulin sensitivity. *Nature Genet.* **20**, 284–287 (1998).
15. Altshuler, D. *et al.* The common PPAR γ Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genet.* **26**, 76–80 (2000).
16. Stefansson, H. *et al.* Neuregulin 1 and susceptibility to schizophrenia. *Am. J. Hum. Genet.* **71**, 877–892 (2002).
17. Van Eerdewegh, P. *et al.* Association of the ADAM33 gene with asthma and bronchial hyperresponsiveness. *Nature* **418**, 426–430 (2002).
18. Gretarsdottir, S. *et al.* The gene encoding phosphodiesterase 4D confers risk of ischemic stroke. *Nature Genet.* **35**, 131–138 (2003).
19. Ozaki, K. *et al.* Functional SNPs in the lymphotoxin- α gene that are associated with susceptibility to myocardial infarction. *Nature Genet.* **32**, 650–654 (2002).
20. Collins, F. S., Guyer, M. S. & Chakravarti, A. Variations on a theme: cataloging human DNA sequence variation. *Science* **278**, 1580–1581 (1997).
21. Kruglyak, L. & Nickerson, D. A. Variation is the spice of life. *Nature Genet.* **27**, 234–236 (2001).
22. Kerem, B. *et al.* Identification of the cystic fibrosis gene: genetic analysis. *Science* **245**, 1073–1080 (1989).
23. Hästbacka, J. *et al.* Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. *Nature Genet.* **2**, 204–211 (1992).
24. Li, W. H. & Sadler, L. A. Low nucleotide diversity in man. *Genetics* **129**, 513–523 (1991).
25. Wang, D. G. *et al.* Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**, 1077–1082 (1998).
26. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
27. Halushka, M. K. *et al.* Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nature Genet.* **22**, 239–247 (1999).
28. Reich, D. E., Gabriel, S. B. & Altshuler, D. Quality and completeness of SNP databases. *Nature Genet.* **33**, 457–458 (2003).
29. Pääbo, S. The mosaic that is our genome. *Nature* **421**, 409–412 (2003).
30. Jeffreys, A. J., Kauppi, L. & Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.* **29**, 217–222 (2001).
31. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
32. Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
33. Abecasis, G. R. *et al.* Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.* **68**, 191–197 (2001).
34. Dawson, E. *et al.* A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418**, 544–548 (2002).
35. Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype structure in the human genome. *Nature Genet.* **29**, 229–232 (2001).
36. Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
37. Johnson, G. C. L. *et al.* Haplotype tagging for the identification of common disease genes. *Nature Genet.* **29**, 233–237 (2001).
38. Carlson, C. S. *et al.* Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nature Genet.* **33**, 518–521 (2003).
39. Goldstein, D. B., Ahmadi, K. R., Weale, M. E. & Wood, N. W. Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. *Trends Genet.* **19**, 615–622 (2003).
40. Taillon-Miller, P. *et al.* Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nature Genet.* **25**, 324–328 (2000).
41. Chakravarti, A. Population genetics—making sense out of sequence. *Nature Genet.* **21**, 56–60 (1999).
42. Chakravarti, A. *et al.* Nonuniform recombination within the human β -globin gene cluster. *Am. J. Hum. Genet.* **36**, 1239–1258 (1984).
43. Tishkoff, S. A. *et al.* Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271**, 1380–1387 (1996).
44. Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton Univ. Press, Princeton, 1994).
45. Foster, M. W. & Sharp, R. R. Race, ethnicity, and genomics: social classifications as proxies of biological heterogeneity. *Genome Res.* **12**, 844–850 (2002).
46. Barbuiani, G., Magagni, A., Minch, E. & Cavalli-Sforza, L. L. An apportionment of human DNA diversity. *Proc. Natl Acad. Sci. USA* **94**, 4516–4519 (1997).
47. Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
48. Jorde, L. B. *et al.* Microsatellite diversity and the demographic history of modern humans. *Proc. Natl Acad. Sci. USA* **94**, 3100–3103 (1997).
49. Dausset, J. *et al.* Centre d'Etude du Polymorphisme Humain (CEPH): collaborative genetic mapping of the human genome. *Genomics* **6**, 575–577 (1990).
50. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
51. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: a fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).
52. The International SNP Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
53. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
54. Lewontin, R. C. The interaction of selection and linkage. I. General considerations: heterotic models. *Genetics* **49**, 49–67 (1964).
55. Hill, W. G. & Robertson, A. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38**, 226–231 (1968).
56. Ohta, T. & Kimura, M. Linkage disequilibrium due to random genetic drift. *Genet. Res.* **13**, 47–55 (1969).
57. Dawson, K. J. The decay of linkage disequilibrium under random union of gametes: how to calculate Bennett's principal components. *Theor. Popul. Biol.* **58**, 1–20 (2000).
58. Langley, C. H. & Crow, J. F. The direction of linkage disequilibrium. *Genetics* **78**, 937–941 (1974).
59. Maniatis, N. *et al.* The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by haplotype analysis. *Proc. Natl Acad. Sci. USA* **99**, 2228–2233 (2002).
60. Hudson, R. R. Estimating the recombination parameter of a finite population model without selection. *Genet. Res.* **50**, 245–250 (1987).
61. Fearnhead, P. & Donnelly, P. Estimating recombination rates from population genetic data. *Genetics* **159**, 1299–1318 (2001).
62. McVean, G., Awadalla, P. & Fearnhead, P. A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**, 1231–1241 (2002).
63. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
64. Clayton, E. W. The complex relationship of genetics, groups, and health: what it means for public health. *J. Law Med. Ethics* **30**, 290–297 (2002).
65. Foster, M. W. & Sharp, R. R. Genetic research and culturally specific risks: one size does not fit all. *Trends Genet.* **16**, 93–95 (2000).
66. Sharp, R. R. & Foster, M. W. Involving study populations in the review of genetic research. *J. Law Med. Ethics* **28**, 41–51 (2000).
67. Marshall, P. A. & Rotimi, C. Ethical challenges in community-based research. *Am. J. Med. Sci.* **322**, 241–245 (2001).
68. Juengst, E. T. Commentary: what “community review” can and cannot do. *J. Law Med. Ethics* **28**, 52–54 (2000).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank many people who contributed to this project: J. Beck, C. Beiswanger, D. Coppock, J. Mintzer and L. Toji at the Coriell Institute for Medical Research for transforming the samples, distributing the DNA and cell lines, and storing the samples for use in future research; J. Greenberg and R. Anderson of the NIH National Institute of General Medical Sciences (NIGMS) for providing funding and support for cell-line transformation and storage in the NIGMS Human Genetic Cell Repository at the Coriell Institute; K. Wakui at Shinshu University for assistance in transforming the Japanese cell lines; N. Carter and D. Willey at the Wellcome Trust Sanger Institute for flow sorting the chromosomes and for library construction, respectively; M. Deschesnes and B. Godard for assistance at the University of Montréal; C. Darmond-Zwaig, J. Olivier and S. Roumy at McGill University and Génome Québec Innovation Centre; C. Allred, B. Gillman, E. Kloss and M. Rieder for help in implementing data flow protocols; S. Olson for work on the website explanations; S. Adeniyi-Jones, D. Burgess, W. Burke, T. Citrin, A. Clark, D. Cowhig, P. Epps, K. Hofman, A. Holt, E. Juengst, B. Keats, J. Levin, R. Myers, A. Obuoforibo, F. Romero, C. Tamura and A. Williamson for providing advice on the project to NIH; A. Peck and J. Witonsky of the National Human Genome Research Institute (NHGRI) for help with project management; E. DeHaut-Combs and S. Saylor of NHGRI for staff support; M. Gray for organizing phone calls and meetings; the people of Tokyo, Japan, the Yoruba people of Ibadan, Nigeria, and the community at Beijing Normal University, who participated in public consultations and community engagements; and the people in these communities who were generous in donating their blood samples. This work was supported in part by Genome Canada, Génome Québec, the Chinese Ministry of Science and Technology, the Chinese Academy of Sciences, the Natural Science Foundation of China, the Hong Kong Innovation and Technology Commission, the University Grants Committee of Hong Kong, the Japanese Ministry of Education, Culture, Sports, Science and Technology, the Wellcome Trust, the SNP Consortium, the US National Institutes of Health (FIC, NCI, NCRR, NEI, NHGRI, NIA, NIAAA, NIAID, NIAMS, NIBIB, NIDA, NIDCD, NIDCR, NIDDK, NIEHS, NIGMS, NIMH, NINDS, OD), the W.M. Keck Foundation and the Delores Dore Eccles Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.B. (drb@sanger.ac.uk) or M.F. (fost1848@msmailhub.oulan.ou.edu).

24, University of Tsukuba, Eubios Ethics Institute, P.O. Box 125, Tsukuba Science City 305-8691, Japan; 25, Howard University, National Human Genome Center, 2216 6th Street NW, Washington, District of Columbia 20059, USA; 26, University of Ibadan College of Medicine, Ibadan, Oyo State, Nigeria; 27, Case Western Reserve University School of Medicine, Department of Bioethics, 10900 Euclid Avenue, Cleveland, Ohio 44106, USA; 28, University of Utah, Eccles Institute of Human Genetics, Department of Human Genetics, 15 North 2030 East, Salt Lake City, Utah 84112, USA; 29, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA; 30, Johns Hopkins University School of Medicine, McKusick–Nathans Institute of Genetic Medicine, 600 North Wolfe Street, Baltimore, Maryland 21287, USA; 31, University of Oxford, Department of Statistics, 1 South Parks Road, Oxford OX1 3TG, UK; 32, University of Oxford, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK; 33, University of Michigan, Department of Biostatistics, Center for Statistical Genetics, 1420 Washington Heights, Ann Arbor, Michigan 48109, USA; 34, North Carolina State University, Bioinformatics Research Center, Campus Box 7566, Raleigh, North Carolina 27695, USA; 35, US National Institutes of Health, National Human Genome Research Institute, 50 South Drive, Bethesda, Maryland 20892, USA; 36, US National Institutes of Health, National Library of Medicine, National Center for Biotechnology Information, 8600 Rockville Pike, Bethesda, Maryland 20894, USA; 37, Chinese Academy of Social Sciences, Center for Applied Ethics, 2121, Building 9, Caoqiao Xinyuan 3 Qu, Beijing 100054, China; 38, Genetic Interest Group, 4D Leroy House, 436 Essex Road, London N1 3QP, UK; 39, Kyoto University, Institute for Research in Humanities, Ushinomiya-cho, Sakyo-ku, Kyoto 606-8501, Japan; 40, Nagasaki University Graduate School of Biomedical Sciences, Department of Human Genetics, Sakamoto 1-12-4, Nagasaki 852-8523, Japan; 41, University of Montréal, The Public Law Research Centre (CRDP), P.O. Box 6128, Downtown Station, Montréal, Québec H3C 3J7, Canada; 42, University of Oklahoma, Department of Anthropology, 455 West Lindsey Street, Norman, Oklahoma 73019, USA; 43, Vanderbilt University, Center for Genetics and Health Policy, 507 Light Hall, Nashville, Tennessee 37232, USA; 44, Wellcome Trust, 183 Euston Road, London NW1 2BE, UK; 45, Washington University School of Medicine, Genome Sequencing Center, Box 8501, 4444 Forest Park Avenue, St Louis, Missouri 63108, USA; 46, Whitehead Institute/MIT Center for Genome Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA; 47, Chinese Academy of Sciences, 52 Sanlihe Road, Beijing 100864, China; 48, Chinese Ministry of Science and Technology, 15B Fuxing Road, Beijing 100862, China; 49, Genome Canada, 150 Metcalfe Street, Suite 2100, Ottawa, Ontario K2P 1P1, Canada; 50, McGill University, Office of Technology Transfer, 3550 University Street, Montréal, Québec H3A 2A7, Canada; 51, Génomique Québec, 630 Boulevard René-Lévesque Ouest, Montréal, Québec H3B 1S6, Canada; 52, Ministry of Education, Culture, Sports, Science and Technology, 3-2-2 Kasumigaseki, Chiyodaku, Tokyo, Japan; 53, Hiraki and Associates, Toranomon No. 5 Mori Building, 17-1, Toranomon 1-Chome, Minato-Ku, Tokyo 105-0001, Japan; 54, The SNP Consortium, 3 Parkway North, Deerfield, Illinois 60015, USA; 55, GlaxoSmithKline, 5 Moore Drive, Research Triangle Park, North Carolina 27709, USA; 56, US National Institutes of Health, National Human Genome Research Institute, 31 Center Drive, Bethesda, Maryland 20892, USA; 57, Foundation for the National Institutes of Health, 1 Cloister Court, Bethesda, Maryland 20892, USA; 58, US National Institutes of Health, Office of Technology Transfer, 6011 Executive Boulevard, Rockville, Maryland 20852, USA; 59, University of Maryland School of Law, 500 West Baltimore Street, Baltimore, Maryland 21201, USA; 60, Herbert Smith, Exchange House, Primrose Street, London EC2A 2HS, UK; 61, Johns Hopkins University School of Medicine, Howard Hughes Medical Institute and the McKusick–Nathans Institute of Genetic Medicine, 725 North Wolfe Street, Baltimore, Maryland 21205, USA; 62, Stanford Center for Biomedical Ethics, 701A Welch Road, Palo Alto, California 94304, USA; 63, New York University, Department of Sociology, 269 Mercer Street, New York, New York 10003, USA; 64, University of California, Berkeley, Department of Sociology, 2420 Bowditch, Berkeley, California 94720, USA; 65, University of New Mexico Health Sciences Center, 214 East Nizhoni Boulevard, Gallup, New Mexico 87301, USA; 66, University of California, Los Angeles, School of Medicine, 695 Charles E. Young Drive South, Los Angeles, California 90095, USA; 67, University of Michigan, Department of Human Genetics, 1241 East Catherine Street, Ann Arbor, Michigan 48109, USA; 68, University of Wisconsin School of Law, 975 Bascom Mall, Madison, Wisconsin 53706, USA; 69, The London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK; 70, Genetic Alliance, 4301 Connecticut Avenue NW, Suite 404, Washington, District of Columbia 20008, USA; 71, University of Washington, Department of Genome Sciences, Box 357730, Seattle, Washington 98125, USA; 72, University of Chicago, Department of Ecology and Evolution, 1101 East 57th Street, Chicago, Illinois 60637, USA; 73, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, USA; 74, US National Institutes of Health, National Human Genome Research Institute, 49 Convent Drive, Bethesda, Maryland 20892, USA

Recent ice ages on Mars

James W. Head¹, John F. Mustard¹, Mikhail A. Kreslavsky^{1,2}, Ralph E. Milliken¹ & David R. Marchant³

¹Department of Geological Sciences, Brown University, Providence, Rhode Island 02912, USA

²Astronomical Institute, Kharkov National University, Kharkov, 61077, Ukraine

³Department of Earth Sciences, Boston University, Boston, Massachusetts 02215, USA

A key pacemaker of ice ages on the Earth is climatic forcing due to variations in planetary orbital parameters. Recent Mars exploration has revealed dusty, water-ice-rich mantling deposits that are layered, metres thick and latitude dependent, occurring in both hemispheres from mid-latitudes to the poles. Here we show evidence that these deposits formed during a geologically recent ice age that occurred from about 2.1 to 0.4 Myr ago. The deposits were emplaced symmetrically down to latitudes of $\sim 30^\circ$ —equivalent to Saudi Arabia and the southern United States on the Earth—in response to the changing stability of water ice and dust during variations in obliquity (the angle between Mars' pole of rotation and the ecliptic plane) reaching $30\text{--}35^\circ$. Mars is at present in an 'interglacial' period, and the ice-rich deposits are undergoing reworking, degradation and retreat in response to the current instability of near-surface ice. Unlike the Earth, martian ice ages are characterized by warmer polar climates and enhanced equatorward transport of atmospheric water and dust to produce widespread smooth deposits down to mid-latitudes.

Of the planets in the Solar System, the climate of Mars is most similar to that of the Earth. Small quasi-periodic variations in the Earth's orbit and spin parameters over timescales of $10^4\text{--}10^5$ years drive large-scale changes in climate^{1,2}. Similarly, Mars exhibits variability of its orbital and axial elements with timescales comparable to the Earth's; however, the ranges of Mars' variations are significantly greater. Over the past 10 Myr, Earth's obliquity has ranged from 22° to 24.5° , while for Mars^{3–6} the range was 14° to 48° . Similarly, the eccentricity of Earth's orbit varied from 0 to 0.06, while for Mars the range was 0 to 0.12. The consequent changes to insolation and seasonality at middle to high latitudes on Mars have inevitably caused significant changes in the seasonal cycles of carbon dioxide, water and dust^{7,8}. Thus Mars may have experienced the most significant quasi-periodic variations in its climate over the past 10 Myr of any planet in the Solar System.

Analysis of data from spacecraft missions before Mars Global Surveyor (MGS) identified laminations and layers in the north and south polar caps whose origin was attributed to quasi-periodic climate change⁹. Theoretical modelling of water vapour diffusion in near surface soils predicted that ground-ice stability in the mid-latitudes would vary in concert with the orbital and axial forcing¹⁰. An array of landforms in the mid-latitudes of Mars were catalogued¹¹ that have analogues to ice-rich terrains on Earth. However, their ages appeared to be significantly greater than 10 Myr, and their origin was not linked to quasi-periodic climate change. Here we use data from the recent Mars missions (MGS and Odyssey) to show multiple lines of evidence for surface deposits that formed as a result of recent quasi-periodic climate change on Mars. The observations span a large range of scales (metres to kilometres to hundreds of kilometres), are of diverse nature (morphology, topography and chemistry), and are remarkably consistent with models of current and past ground-ice stability. These results all point to the presence of a succession of metres-thick, latitude-dependent surface deposits that are young, ice-rich when formed, and whose deposition and removal is driven by climate change that is induced by spin-axis tilting.

MGS observations

The subkilometre-scale roughness derived from global Mars Orbiter Laser Altimeter (MOLA) data revealed smoothing at subkilometre scales above about 30° latitude in both hemispheres¹² (Fig. 1a). This topographic smoothing was attributed at least partly to a surface deposit several metres thick that was superposed, or draped on, older geological units (Fig. 2) and referred to as a mantle^{12–14}. The

statistics of the kilometre- and subkilometre-scale topography of Mars are influenced by such a metres-thick mantle because the typical topographic slopes at these scales are of the order of a degree or less; this means that the typical vertical scale of topography is two orders of magnitude shorter than the spatial scale. Another prominent latitudinal trend is seen in topographic concavity at hundreds of metres baseline (Figs 1b and 3c), a measure of the relative balance of concave and convex segments of topographic profiles¹³.

Analysis of Mars Orbiter Camera (MOC) images revealed the distribution of many geological features that also exhibit a latitude dependence^{15–23}. A systematic study of the presence or absence of a specific morphology representing a metres-thick but partially degraded and discontinuous surface deposit (Fig. 2c–f) showed that this terrain was limited to the $30^\circ\text{--}60^\circ$ north and south latitude zone (Figs 1c and 3b)¹⁶. This deposit was interpreted as a recent, formerly ice-rich dust mantle that originated as a thin blanketing airfall layer and was now undergoing dissection and removal¹⁶. There was no evidence of this deposit in equatorial and low-latitude regions (between 30°N and 30°S)¹⁶. Poleward of 60° , this mantle was continuous and commonly characterized by a unique bumpy texture at the scale of several tens of metres that often resembled the surface of a basketball (Fig. 2a), and lineated and wrinkled versions of this terrain type^{12–16,22,23}. Evidence for multiple layers (Fig. 2b) within the mantle unit has been documented, indicating several generations of deposition and removal^{13,20}.

Different types of polygons and patterned ground were also noted to occur preferentially above 30° north and south latitude^{17,18}, suggesting the presence of a dynamic layer in which near-surface ice-rich material was undergoing thermal cycling. Startling images show what appear to be very recent water-carved gullies that occur preferentially in this region and have been interpreted to represent groundwater sapping¹⁹, melting of ground ice²⁴ or snowpack during higher obliquity²⁵, or local microclimate conditions²⁶. Viscous flow features²⁰ and related deposits²¹ suggestive of mobilization and flow of ice-rich mantling materials occur in the same latitude band. Results from the γ -ray and neutron spectrometers on board Mars Odyssey^{27–29} showed the distribution of hydrogen within the top metre of the martian surface. The high concentrations of hydrogen at polar latitudes were interpreted to indicate the presence of water ice (at $>50\%$ by mass) in the near subsurface, below a hydrogen-free surface sediment layer tens of centimetres thick³⁰. The global distribution of hydrogen and interpreted water abundance³⁰ on Mars (Fig. 1e) shows a remarkable correlation with the latitude-dependent deposits interpreted from the MOLA and MOC data³¹.

Furthermore, it shows a very high correlation with the present stability of near-surface ground ice (Fig. 1d) as modelled from the diffusive exchange of water vapour between the regolith pore space and the martian atmosphere^{10,32}.

Latitude is the single variable with which all of these diverse observations correlate, and climate is the only process known to be latitude-dependent. Degradation of the youthful mantle in the mid-latitudes suggests a very recent change in climate¹⁶. The remarkable correspondence (Fig. 1) between the character of the terrain smoothness, the continuity of the mantle, the high hydrogen

abundance, and the theoretical stability of ice in the near-surface soil provide compelling evidence for climate-driven water ice and dust mobility, emplacement and dissection. What do the characteristics of these regions and deposits tell us about the nature and timing of such activity?

Interpretations of deposits

One explanation for the correlation between the observations and models is that we are simply observing an outcome of climate-related diffusive exchange of water vapour between the regolith pore

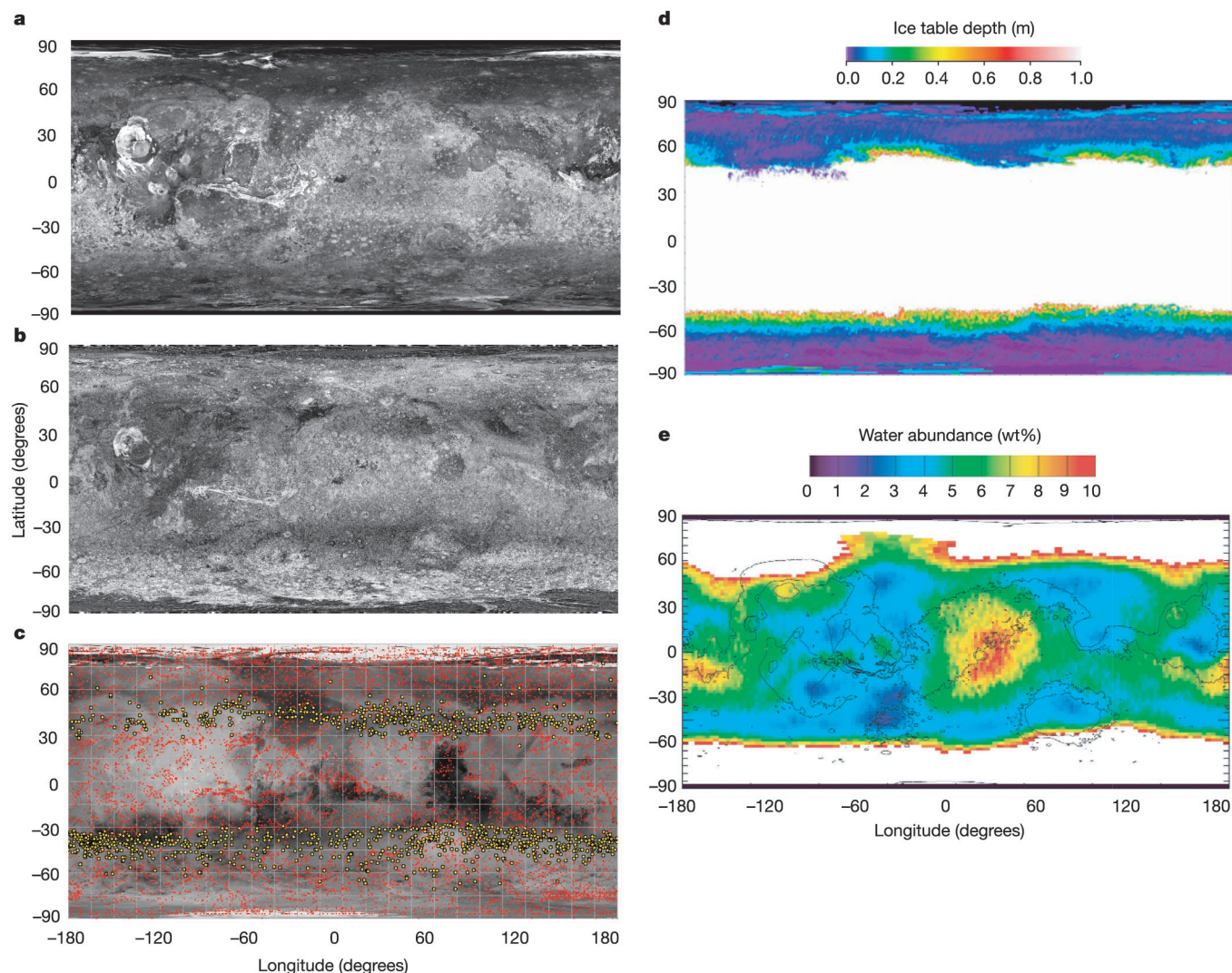


Figure 1 Maps showing important latitudinal trends on Mars (simple cylindrical projection). **a**, Subkilometre-scale topographic roughness. Brighter shades denote a rougher surface. The roughness parameter mapped is the interquartile width of the frequency distribution of the differential slope¹² at 0.6 km baseline. **b**, Subkilometre-scale topographic concavity. Brighter shades denote higher prevalence of concave topography. The parameter mapped is the median curvature of topographic profiles at 0.6 km baseline normalized by the interquartile width of the curvature–frequency distribution¹³. Most of the geologic units in the equatorial zone have some concave topography (positive concavity). The high-latitude zones have higher positive concavity (Fig. 3c). Smooth filling of local lows by the mantle here eliminates smaller-scale topographic variations and in this way amplifies the concavity. The mid-latitude zones (30°–60°) exhibit a distinctive low concavity that differs from both unmantled equatorial and high-latitude mantled regions¹³. This is interpreted to be related to dissection of the mantle^{13–16}, small-scale irregular topographic variations due to dissected and patchy mantle occurrence in these zones cause randomly alternating concave and convex profile segments and concavity values close to zero. All data in **a** and **b** are derived from statistics of all along-track

topographic profiles obtained by MOLA. **c**, Distribution of examined MOC images. Yellow circles indicate MOC images with dissected mantle terrain, red circles indicate images with no apparent dissected terrain^{16,20}. The mantle is interpreted to be present poleward of ~60°, but is not dissected. An albedo mosaic is used as a background. **d**, Map of ice table depth for an annual mean of 10 precipitable micrometres of atmospheric water vapour based on a near-surface ice stability model³². In this model, ice in the near subsurface is unstable at low latitudes (equatorial regions, no ice-rich mantle terrains observed) and stable at high latitudes (polar regions; thick, intact ice-rich mantles observed). The mid-latitude regions are transitional zones in which ice is currently unstable, corresponding to areas where we observe degraded, previously ice-rich deposits such as the dissected mantle terrain. This model of ice stability shows a remarkable correlation with the observations of ice-rich deposits from Odyssey^{27–30} and MGS data. **e**, Mars Odyssey GRS/NS experiment data showing interpreted water abundance in weight per cent³⁰. White areas above ~60° latitude are interpreted as ice buried a few centimetres below the surface³⁰.

space and the martian atmosphere. The depth to ice below the surface is predicted to have varied vertically in the regolith and with latitude as Mars experienced various orbital cycles in its most recent history¹⁰. In this scenario, the latitude-dependent surface textures in the 30°–60° zones would be due to reworking of the surface caused by repeated diffusive precipitation and subsequent sublimation of water ice during recent excursions in ground-ice stability. The 30°–60° zones would thus represent a zone of dynamic exchange between regolith and polar water reservoirs, while regions poleward of 60° represent a remnant stable deposit of water ice in the regolith under current conditions. In this scenario^{10,32}, no significant amounts of new material would be deposited on the surface.

There is, however, abundant evidence for deposition of dust–ice mixtures and layering (Fig. 2). Superposition relations are common¹⁶, dissection in the 30°–50° latitude range shows removal and exposure of underlying units¹⁶, and multiple layers are seen,

particularly toward the poles¹³. The latitudinal trend of subkilometre-scale roughness is readily explained by the deposition of a metres-thick smooth layer on top of underlying terrain¹². Such a deposit explains the observed increase of topographic concavity above 60° latitude¹³, while the decrease in concavity in the 30°–60° zones is interpreted to be related to dissection of the mantle^{13,16}. The extreme morphological homogeneity of the unit above 60° latitude and the manner in which it drapes subjacent deposits²² also supports deposition of a mantle. The very low abundance of superposed impact craters^{13,16,23} supports a very youthful age (<10 Myr), and the paucity of subkilometre-scale degraded impact craters²³ supports deposition, mantling and obscuration, rather than *in situ* reworking. The very high near-surface water-ice content inferred from the Odyssey γ -ray and neutron spectrometer data^{27–30} is consistent with a model of deposition of a relatively uniform ice-rich dust layer from the atmosphere. Other evidence for the ice-rich nature of the deposit includes the correspondence of its continuous extent with modelled ground-ice stability^{10,32}, its apparent mechanical strength²⁰, evidence for creep on steep slopes^{20,39}, and its relation to water and ice-related features (such as viscous flow, and polygons)^{19–21,24,25}. The weight of the geological observations points to airfall deposition.

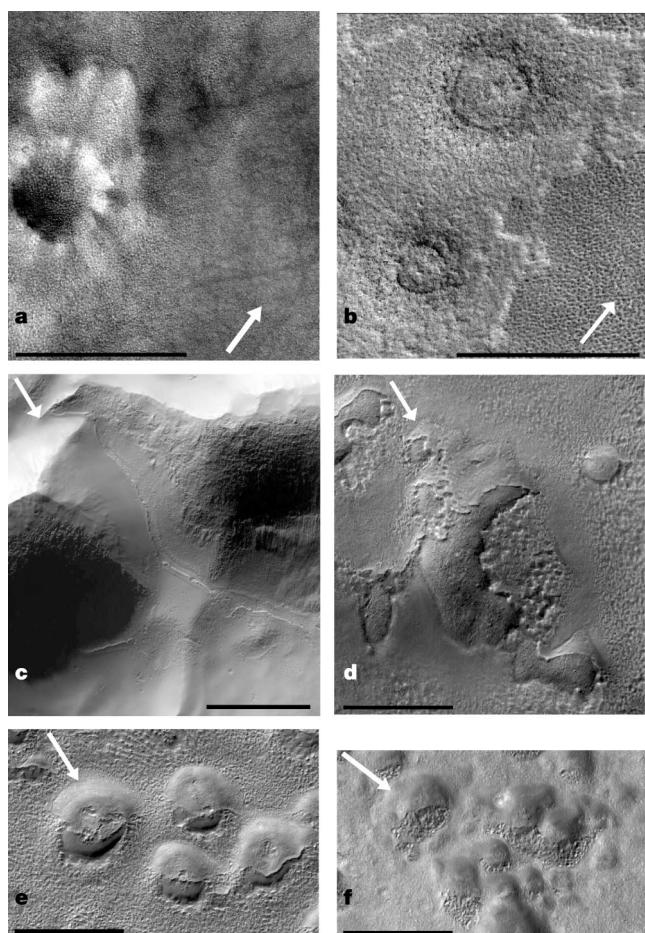


Figure 2 Characteristic tens of metre-scale textures of the mantle and dissected terrain as seen in MOC high-resolution images^{15,16}. Scale bars at the bottom of each image are 1 km; north is at the top. Arrows show the illumination direction. **a**, Homogeneous mantle with basketball texture covers well-preserved crater and its ejecta. Portion of MOC image E02-01380; 61° N, 210° W. **b**, A mantle layer with the basketball texture partly removed. Co-registered MOLA profile shows a 3–4 m height for the topographic step across the mantle layer boundary. Two circular features are probably old, heavily degraded, mantled impact craters. Portion of MOC image M02-01316; 69° N, 150° W. **c**, Example of dissected mantle terrain with internal layering found in the mid-latitudes. The mantle is smooth and has been removed from the pole-facing slopes. The removal of the mantle on the shaded side of the rightmost knob has revealed an underlying mantle layer. Portion of MOC image M20-00144; 35° S, 174° W. **d–f**, Examples of dissected mantle found in the mid-latitude regions where the mantle has been preferentially removed from pole-facing slopes. **d**, Portion of MOC image M04-02856; 47° S, 218° W. **e**, Portion of MOC image FHA-01450; 44° S, 240° W. **f**, Portion of MOC image M04-02289; 43° S, 240° W.

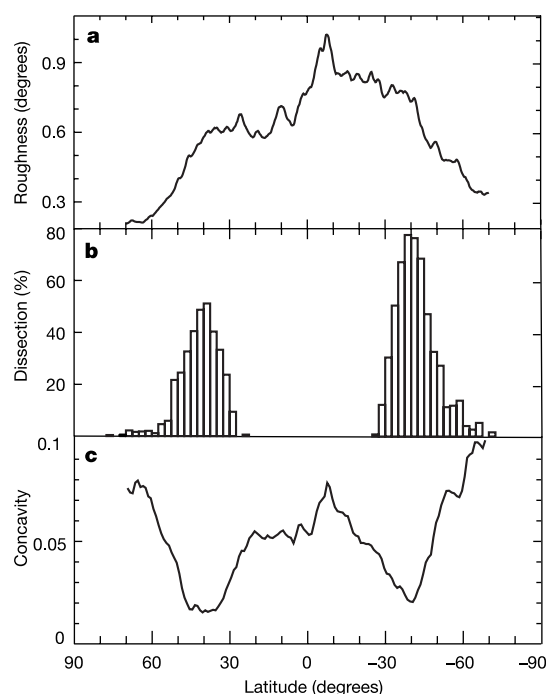


Figure 3 Latitudinal dependence of statistical characteristics of subkilometre-scale surface topography (roughness and concavity) and percentage of images with dissected terrain. Southern latitudes are shown as negative values. **a**, Roughness parameter plotted is the interquartile width of the frequency distribution of the differential slope¹² at 0.6 km baseline. The peak at 10° S is related to the Valles Marineris complex; the difference in roughness between 20–40° S and 20–40° N reflects the geological dichotomy of Mars, with prevalence of rough heavily cratered highlands in the southern hemisphere and smoother lowlands in the northern hemisphere. The decrease of roughness poleward from 40° N and 40° S does not correlate with large geomorphological features and reflects the latitudinal smoothing trend. **b**, Percentage of high-resolution MOC images having the characteristic dissected morphology among 15,000 images surveyed¹⁶, in 2.5° latitude bins. There are two peaks in the presence of the dissected terrain at 41° N and 41° S, which correspond to the latitude in **a** where smoothing begins. **c**, Topographic concavity. The parameter plotted is the median curvature of topographic profiles at 0.6 km baseline normalized by the interquartile width of the curvature–frequency distribution¹³. A strong decrease of concavity is observed within the 30–50° zone in both hemispheres and correlates well with the occurrence of dissected terrains.

Integration of observations and climate models

We propose a model that explains the observations by the deposition and removal of mixtures of dust and water ice, and is controlled by climate variations resulting from quasi-periodic variations in orbital parameters. During periods of high obliquity ($>30^\circ$), increased summer insolation on the polar regions results in greater water vapour release from the polar caps, increasing the atmospheric water vapour content and thus the humidity^{33,34}. Higher humidity moves the latitude of subsurface ice stability closer to the equator, and within the soil profile, closer to the surface¹⁰. Climate models predict that ice removed from the poles is deposited on the surface at these lower latitudes where it persists throughout the year^{35–38}. Mars global circulation model (GCM) simulations also predict strong winds favourable for dust lifting from global reservoirs and redistribution during such times^{35,36}. This dust is readily incorporated into the growing ice deposits. These three factors (increased atmospheric water vapour content, increased atmospheric dust content, and shifting of the surface ice stability zone toward the equator) are thus predicted to act in concert at high obliquity to produce the latitude-dependent deposits that we observe. Owing to the higher year-average and summer-average polar temperatures and denser atmosphere, the transport of water at high obliquities is predicted to be rapid³⁷. High obliquity periods favourable to this process are of the order of 40 kyr (Fig. 4)⁵, long enough to provide deposition of at least several metres of ice-rich material³⁷. Thus, in our model the deposited layers should be geologically recent, latitude-dependent, dust–ice mixtures, superposed on underlying terrain, and varying in their character as a function of latitude.

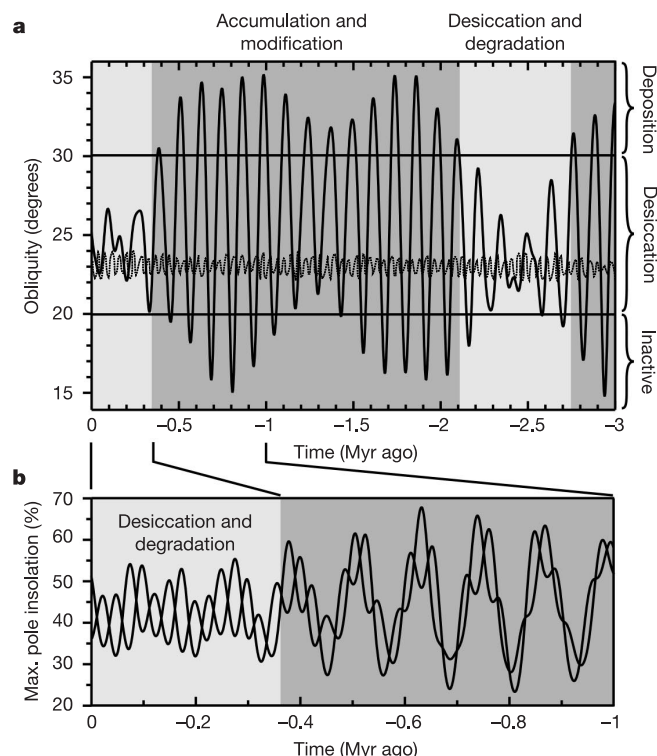


Figure 4 Orbital forcing of climate in the past. **a**, Obliquity variations⁵ for the past 3 Myr, with glacial (accumulation and modification; dark grey) and interglacial (desiccation and degradation; pale grey) periods marked. Low-amplitude line between 22° and 24° represents the obliquity range on Earth during the comparable period of history⁵. **b**, Maximal insolation of the north and south poles⁵ for the past 1 Myr. During the past 300 kyr, there is an asymmetry in insolation, but before 300 kyr ago, insolation is primarily driven by obliquity and is symmetric.

As obliquity decreases, the surface ice stability zones shrink toward the polar caps and ice sublimates from the uppermost layer of the mantle, creating a protective ice-poor crust analogous to the sublimation of ice from loess on Earth³⁹. At lower obliquity, water remains sequestered in cold polar regions and the atmospheric humidity is decreased. Under this drier atmosphere, calculations suggest poleward shrinking of this ice-rich mantle and widespread desiccation¹⁰ between 30° and 60° . As the ice sublimates, it leaves behind a weakly cemented, porous surface lag (analogous to the lags observed in the Fox permafrost tunnel³⁹) that retards sublimation. If desiccation is severe and the remaining ice-poor regolith lacks a cohesive surface crust, then deposits will become susceptible to eolian disruption and destruction¹⁶. For high latitudes ($>60^\circ$), ground ice is stable near the surface with obliquity as low as 20° , and thus over the past 10 Myr dissection would be most important in the mid-latitudes (30° – 60°). The depth of desiccation is predicted to vary as the square root of time³⁹, and thus the characteristic timescale of water loss increases with time and decreasing obliquity. Thus, low-obliquity periods are likely to be too short to desiccate and remove the mantle completely at all latitudes. Evidence for the present desiccation and removal of the mantle layers includes the dissected terrain in the 30° – 60° north and south latitude zone^{16,20} (Fig. 2e, f, and Fig. 3), the decrease of subkilometre-scale topographic concavity (Figs 1b and 3c) in this same latitude range¹³, and the correspondence of this zone with the latitudes predicted to be at present desiccated (Fig. 1e, f) in the near surface^{10,32}.

A variety of surface features at higher latitudes ($>60^\circ$) (for example, basketball-textured terrain and its variants, and polygons) are interpreted to represent the net effect of long-term orbital variations on relatively stable, debris-covered ice-rich deposits. In the Antarctic dry valleys, a cold-polar desert that represents a terrestrial analogue for regions with buried ice on Mars, such landforms are related to sublimation and thermal cycling of ice-rich sediments^{40,41}. In Beacon Valley, sublimation of debris-rich ice produces a dry surface lag that insulates and slows loss of remaining ice⁴¹. Ground temperature cycling between $<0^\circ\text{C}$ and $\ll 0^\circ\text{C}$ creates tensile stresses that result in a network of hexagonal cracks, extending upward from buried ice to the ground surface (and vice versa). A lag of coarse-grained sublimation till forms on top of the ice, where fines sift into open thermal contraction cracks. Owing to these spatial variations in till texture, rates of sublimation vary across the ice surface. High rates occur below coarse-grained lags that cap contraction cracks, and low rates are found at polygon centres beneath fine-grained debris of low porosity and permeability. The high-centred 'sublimation' polygons of Beacon Valley may be analogous to the latitude-dependent basketball-textured terrain on Mars. If so, the 'basketball' terrain probably reflects a combination of long-term ice sublimation, coupled with thermal cycling on timescales both short (seasonal) and long (for example, obliquity)⁴¹. A close analogue for such modification on Earth includes long-term changes in the distribution and character of permafrost regions of the high Arctic^{42,43} and Antarctic.

Estimates of the age of the mantling unit and its contained layers come from several sources. The number of small fresh craters is very low, and the crater retention age is correspondingly very young¹⁶, possibly as young as 150 kyr but most certainly less than 10 Myr. Stratigraphic relations show a complex history of successive episodes of deposition and removal¹³, suggesting that the mantle consists of multiple depositional cycles. Regional stratigraphic evidence at high latitudes^{9,14} shows outliers of polar cap deposits as well as dark dunes locally superposed on the mantle. Finally, many of the surface textures of the mantle (for example, polygons, and basketball-textured terrain) are consistent with morphologies formed over several glacial cycles in the Arctic and Antarctic periglacial environments, suggesting that changing environmental conditions over extended time periods are required. Thus, taken together, the characteristics of the mantle suggest that the deposit is

geologically very young, but that it is composed of numerous sublayers and features requiring multiple depositional cycles.

Integrating the geologic evidence with the most recent orbital history of Mars results in a dynamic depositional–erosional model for ice in the mid–high latitudes (Fig. 4). The orbital and spin elements of obliquity, eccentricity, and the areocentric longitude of the Sun at perihelion (L_p) (Fig. 4) all affect water-ice migration on Mars^{10,37}, but with varying timescales and degree. Over the past 300 kyr the obliquity has remained relatively stable within a range of 22°–26° (Fig. 4). During this period, high eccentricity causes differences in seasonality of the northern and southern hemispheres (for example, the southern hemisphere today experiences shorter, warmer summers and longer winters than the northern hemisphere), while retarding of L_p causes the seasonality difference to reverse every ~25 kyr. If the retardation of L_p were the dominant driver of the deposition and erosion of the mantle, then we would expect the geologic evidence to be asymmetric in its preservation, reflecting the last such state (Fig. 4b). On a global scale, however, the geologic evidence is remarkably symmetric (Figs 1 and 3). But earlier than 300 kyr ago, obliquity regularly exceeded 30°, totalling 15 discrete excursions over the past 2 Myr, during which each excursion lasted of the order of 20–40 kyr. During this period, the high-amplitude obliquity oscillations dominate the insolation regime of the polar regions. When obliquity exceeds 30°, water ice is stable in the near surface down to the lower mid-latitudes¹⁰, and climate models predict that water ice is efficiently moved from polar cap reservoirs by sublimation and deposited in the mid-latitudes³⁷. According to Mars climate models, each high-obliquity excursion would result in tens of metres of ice being removed from the poles; this material would then be transported, and deposited in the mid-latitudes with a water ice content equivalent to a few metres³⁷, together with dust also eroded from the polar layered terrains⁹.

As the obliquity changes to values <30°, ice is sublimated from the mid-latitude deposits, creating a surface lag of dust analogous to that seen in Earth environments^{39,41}. Such a lag deposit strongly retards further sublimation³⁹. The lower year-average and summer temperatures at high latitudes at low obliquity decrease the diffusion rate of water vapour in the pores. Furthermore, as obliquity declines below 21°, the atmospheric pressure drops significantly with the sequestering of CO₂ in the polar caps⁴⁴, further decreasing the diffusion rate in small pores and preventing eolian erosion of the lag. The effects of a retarding loess-like lag allows for all or part of the mantle deposited during a high-obliquity period to be preserved during the low-obliquity excursion. It is during the extended periods of moderate obliquity, such as for the past 300 kyr, that geologic processes have sufficient time to produce large-scale degradation of the mantle in mid-latitude zones.

Before 5 Myr ago, martian obliquity regularly exceeded 45° and the mean obliquity over this time was >35°, compared to a mean of 26° for the past 4 Myr (ref. 5). With obliquity >45° (>6 Myr ago), the polar sublimation rate is very high and ice is predicted to persist at the equator³⁵. Given the relatively sharp boundary of the mantle at mid-latitudes and the lack of evidence of an extensive remnant mantle in equatorial regions¹⁶, we consider this time period a less likely option for formation of the present-day high-latitude mantle. If, however, mantle deposition and preservation is sustained by the mean obliquity, rather than the excursions, then the period before 5 Myr ago may be a more likely candidate.

Ice ages on Mars

Integrating the observations with global climate models results in a martian ice age hypothesis (Fig. 4a). Between about 2,100 and 400 kyr ago, obliquity regularly exceeded 30° and water ice was removed from the polar reservoirs and transported to mid-latitudes, where it nucleated on dust particles in the air and/or at the surface and was deposited as a mantle. Though obliquity varied greatly during this time, the periods with lower obliquity were too short to allow for

complete erosion of the sublimation-retarding surface lags, preserving much of the mantle intact. We refer to such periods of net deposition (for example, between 2,100 and 400 kyr ago) as glacial periods. During the past ~300 kyr, obliquity has been near 25° with comparatively little variation. We refer to these times as interglacial periods (Fig. 4a). Water ice in the mid-latitudes has been slowly and steadily removed from this reservoir by diffusion, sublimation, and atmospheric transport processes; it was deposited in the polar regions, creating the uppermost layers in the polar cap^{45,46}, and also resulting in the degradation of the surface mantle in the 30°–60° latitude bands. Because of the thinness of the deposit and the presence of admixed dust, the total volume of water is of the order of a metre global equivalent layer, and thus only a few per cent of the present volume of the polar caps. If the dissected portion of the deposit (30°–60° latitudes) contained ~50% by volume water ice, and half of this was removed by desiccation processes and deposited at the poles, it would form a layer several tens of metres thick on the polar caps.

We thus conclude that the emplacement of the ice- and dust-rich mantle extending from polar regions down to low mid-latitudes represents the equivalent of an ice age on Mars. If such an ice cover had occurred on Earth, it would have reached southward to latitudes equivalent to Saudi Arabia, North Africa and the southern United States. Poleward migration of the equatorward limit of the ice–dust-rich mantle, partial desiccation of the mantle in the latitudinal belt between 30° and 60°, and modification via sublimation and thermal cycling of long-lived ice-rich deposits poleward of 60°, represent some of the changes that occur during interglacial times (Fig. 4).

Martian ice ages as proposed here differ considerably from those on Earth^{7,47}. First, terrestrial oceans are large heat reservoirs and distribution systems, and sinks for dust. Second, the major reservoir response during climate cycles on Earth is the amount of water stored in the polar caps; the major atmospheric gas components are not thought to change significantly. Third, orbital forcing factors thought to be the pacemaker for the timing of Earth's recent ice ages⁴⁸ are much less extreme than on Mars⁶ (Fig. 4a) owing to the stabilizing presence of the Earth's Moon. On Earth, lower polar insolation causes the deposition of snow, and the formation and lateral spreading of continental ice sheets.

Mars, on the other hand, is characterized by a lack of oceans in its recent past, an abundant and mobile dust supply, and extreme orbital forcing factors. Its major atmospheric gas (CO₂) is in dynamic equilibrium with its solid phase, resulting in the potential for significant changes in atmospheric CO₂ abundance and pressure⁴⁹. Glacial periods on Earth are characterized by colder temperatures at the poles on average, while on Mars, the reverse is true.

In cold polar deserts on Earth, evidence has been cited for the preservation of glacial ice >8 Myr old at depths just below the surface⁵⁰. Mars is characterized by an extremely cold polar desert environment, and we anticipate that the record of recent climate change on Mars is similarly preserved in the shallow subsurface over wide latitude expanses—and thus readily accessible to future robotic and human exploration. □

Received 16 May; accepted 8 October 2003; doi:10.1038/nature02114.

1. Imbrie, J. & Imbrie, K. P. *Ice Ages: Solving the Mystery* (Harvard Univ. Press, Cambridge, MA, 1986).
2. Zachos, J. *et al.* Trends, rhythms, and aberrations in global climate 65 Ma to present. *Science* **292**, 686–693 (2001).
3. Touma, J. & Wisdom, J. The chaotic obliquity of Mars. *Science* **259**, 1294–1297 (1993).
4. Laskar, J. & Robutel, P. The chaotic obliquity of the planets. *Nature* **362**, 608–612 (1993).
5. Laskar, J. *et al.* Orbital forcing of the martian polar layered deposits. *Nature* **419**, 375–377 (2002).
6. Ward, W. R. in *Mars* (eds Kieffer, H. H. *et al.*) 298–320 (Univ. Arizona Press, Tucson, AZ, 1992).
7. Kieffer, H. H. & Zent, A. P. in *Mars* (eds Kieffer, H. H. *et al.*) 1180–1218 (Univ. Arizona Press, Tucson, AZ, 1992).
8. Fanale, F. P. *et al.* in *Mars* (eds Kieffer, H. H. *et al.*) 1135–1179 (Univ. Arizona Press, Tucson, AZ, 1992).
9. Thomas, P. *et al.* in *Mars* (eds Kieffer, H. H. *et al.*) 767–795 (Univ. Arizona Press, Tucson, AZ, 1992).
10. Mellon, M. T. & Jakosky, B. M. The distribution and behavior of Martian ground ice during past and present epochs. *J. Geophys. Res.* **100**, 11781–11799 (1995).
11. Squyres, S. W. *et al.* in *Mars* (eds Kieffer, H. H. *et al.*) 523–554 (Univ. Arizona Press, Tucson, AZ, 1992).

12. Kreslavsky, M. A. & Head, J. W. Kilometre-scale roughness of Mars: Results from MOLA data analysis. *J. Geophys. Res.* **105**, 26695–26711 (2000).
13. Kreslavsky, M. A. & Head, J. W. Mars: Nature and evolution of young latitude-dependent water-ice-rich mantle. *Geophys. Res. Lett.* **29**, doi:10.1029/2002GL015392 (2002).
14. Kreslavsky, M. A. & Head, J. W. Stratigraphy of young deposits in the northern circumpolar region, Mars. *Lunar Planet. Sci.* **XXXIV**, abstr. 1476 (2003).
15. Malin, M. C. & Edgett, K. S. Mars Global Surveyor Mars Orbiter Camera: Interplanetary cruise through primary mission. *J. Geophys. Res.* **106**, 23429–23570 (2001).
16. Mustard, J. F. *et al.* Evidence for recent climate change on Mars from the identification of youthful near-surface ground ice. *Nature* **412**, 411–414 (2001).
17. Seibert, N. M. & Kargel, J. S. Small-scale martian polygonal terrain: Implications for liquid surface water. *Geophys. Res. Lett.* **28**, 899–903 (2001).
18. Mangold, M. *et al.* High latitude patterned grounds on Mars: Evidence for recent melting of near-surface ground ice. *Lunar Planet. Sci.* **XXXIII**, abstr. 1219 (2002).
19. Malin, M. C. & Edgett, K. S. Evidence for recent groundwater seepage and surface runoff on Mars. *Science* **288**, 2330–2335 (2000).
20. Milliken, R. E. *et al.* Viscous flow features on the surface of Mars: Observations from high-resolution MOC images. *J. Geophys. Res.* **108**, doi:10.1029/2002JE002005 (2003).
21. Cabrol, N. A. & Grin, E. A. The recent Mars global warming (MGW) and/or south pole advance (SPA) hypothesis: Global geological evidence and reasons why gullies could still be forming today. *Lunar Planet. Sci.* **XXXIII**, abstr. 1058 (2002).
22. Carr, M. H. Mars Global Surveyor observations of Martian fretted terrain. *J. Geophys. Res.* **106**, 23571–23593 (2001).
23. Kostama, V.-P. *et al.* Morphology of the northern plains in the circumpolar region, Mars. *Lunar Planet. Sci.* **XXXIV**, abstr. 1340 (2003).
24. Costard, F. *et al.* Formation of recent martian debris flows by melting of near-surface ground ice at high obliquity. *Science* **295**, 110–113 (2002).
25. Christensen, P. R. Formation of recent martian gullies through melting of extensive water-rich snow deposits. *Nature* **422**, 45–48 (2003).
26. Hecht, M. H. Metastability of liquid water on Mars. *Icarus* **156**, 373–386 (2002).
27. Boynton, W. V. *et al.* Distribution of hydrogen in the near-surface of Mars: Evidence for subsurface ice deposits. *Science* **296**, 81–85 (2002).
28. Feldman, W. C. *et al.* Global distribution of neutrons from Mars: Results from Mars Odyssey. *Science* **297**, 75–78 (2002).
29. Mitrofanov, I. *et al.* Maps of subsurface hydrogen from the high energy neutron detector, Mars Odyssey. *Science* **297**, 78–81 (2002).
30. Feldman, W. C. *et al.* The global distribution of near-surface hydrogen on Mars. *6th Int. Mars Conf.* Abstr. 5218 (2003).
31. Tokar, R. L. *et al.* Ice concentration and distribution near the south pole of Mars: Synthesis of odyssey and global surveyor analyses. *Geophys. Res. Lett.* **29**, doi:10.1029/2002GL015691 (2002).
32. Mellon, M. T. Theory of ground ice on Mars and implications to the neutron leakage flux. *Lunar Planet. Sci.* **XXXIV**, abstr. 1916 (2003).
33. Jakosky, B. M. & Carr, M. H. Possible precipitation of ice at low latitudes of Mars during periods of high obliquity. *Nature* **315**, 559–561 (1985).
34. Jakosky, B. M., Henderson, B. G. & Mellon, M. T. Chaotic obliquity and the nature of the martian climate. *J. Geophys. Res.* **100**, 1579–1584 (1995).
35. Richardson, M. I. & Wilson, R. J. Investigation of the nature and stability of the Martian seasonal water cycle with a general circulation model. *J. Geophys. Res.* **107**, doi:10.1029/2001JE001536 (2002).
36. Haberle, R. A. *et al.* Orbital change experiments with a Mars general circulation model. *Icarus* **161**, 66–89 (2003).
37. Mischna, M. *et al.* On the orbital forcing of Martian water and CO₂ cycles: A general circulation model study with simplified volatile schemes. *J. Geophys. Res.* **108**, doi:10.1029/2003JE002051 (2003).
38. Richardson, M. I. *et al.* Obliquity, ice sheets, and layered sediments on Mars: What spacecraft observations and climate models are telling us. *Lunar Planet. Sci.* **XXXIV**, abstr. 1281 (2003).
39. Johnson, J. J. & Lorenz, R. D. Thermophysical properties of Alaskan loess: An analog material for the martian polar layered terrain? *Geophys. Res. Lett.* **27**, 2769–2772 (2000).
40. Bockheim, J. G. & Hall, K. J. Permafrost, active-layer dynamics and periglacial environments of continental Antarctica. *S. Afr. J. Sci.* **98**, 82–90 (2002).
41. Marchant, D. R. *et al.* Formation of patterned ground and sublimation till over Miocene glacier ice in Beacon Valley, southern Victoria Land, Antarctica. *Geol. Soc. Am. Bull.* **114**, 718–730 (2002).
42. Williams, P. J. & Smith, M. W. *The Frozen Earth* (Aldon, Oxford, UK, 1989).
43. Dixon, J. C. & Abrahams, A. D. *Periglacial Geomorphology* (Binghampton Symp. in Geomorphology: Int. Ser. No. 22, Wiley and Sons, Chichester, 1992).
44. Fanale, F. P. & Salvail, J. R. Quasi-periodic atmosphere-regolith-cap CO₂ redistribution in the Martian past. *Icarus* **111**, 305–316 (1994).
45. Tanaka, K. L. & Scott, D. H. *Geologic map of the Polar Regions of Mars* (Map I-1802-C, Misc. Invest. Ser., US Geological Survey, Reston, VA, 1987).
46. Kolb, E. & Tanaka, K. Geologic history of the polar regions of Mars based on Mars global surveyor data—II. Amazonian period. *Icarus* **154**, 22–39 (2001).
47. Christie-Blick, N. Pre-Pleistocene glaciation on Earth: Implications for climatic history on Mars. *Icarus* **50**, 408–422 (1982).
48. Imbrie, J. Astronomical theory of the Pleistocene ice ages: A brief historical review. *Icarus* **50**, 408–422 (1982).
49. Leovy, C. Weather and climate on Mars. *Nature* **412**, 245–249 (2001).
50. Sugden, D. E. *et al.* Preservation of Miocene glacier ice in East Antarctica. *Nature* **376**, 412–414 (1995).

Acknowledgements We acknowledge the many discussions that we had with individuals during the 37th Brown-Vernadsky Microsymposium, held at the Lunar and Planetary Institute on 15–16 March 2003. We also thank J. Dixon, A. Côté and P. Neivert for assistance with manuscript preparation. This work was supported by NASA (J.W.H., J.E.M. and M.A.K.) and the NSF, Polar Programs (D.R.M.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.W.H. (James_Head@brown.edu).

Molecular machinery for non-vesicular trafficking of ceramide

Kentaro Hanada, Keigo Kumagai, Satoshi Yasuda, Yukiko Miura, Miyuki Kawano, Masayoshi Fukasawa & Masahiro Nishijima

Department of Biochemistry and Cell Biology, National Institute of Infectious Diseases, 1-23-1, Toyama, Shinjuku-ku, Tokyo 162-8640, Japan

Synthesis and sorting of lipids are essential for membrane biogenesis; however, the mechanisms underlying the transport of membrane lipids remain little understood. Ceramide is synthesized at the endoplasmic reticulum and translocated to the Golgi compartment for conversion to sphingomyelin. The main pathway of ceramide transport to the Golgi is genetically impaired in a mammalian mutant cell line, LY-A. Here we identify CERT as the factor defective in LY-A cells. CERT, which is identical to a splicing variant of Goodpasture antigen-binding protein, is a cytoplasmic protein with a phosphatidylinositol-4-monophosphate-binding (PtdIns4P) domain and a putative domain for catalysing lipid transfer. *In vitro* assays show that this lipid-transfer-catalysing domain specifically extracts ceramide from phospholipid bilayers. CERT expressed in LY-A cells has an amino acid substitution that destroys its PtdIns4P-binding activity, thereby impairing its Golgi-targeting function. We conclude that CERT mediates the intracellular trafficking of ceramide in a non-vesicular manner.

Various steps in lipid biosynthesis occur in different intracellular compartments. In addition, intracellular translocation of lipids may be crucial for lipid-mediated signalling. For these, the intracellular transport of lipids from the sites of their synthesis to their appropriate destinations must occur. The trafficking of integral membrane proteins in eukaryotic cells is mediated by transport vesicles, which load the desired set of proteins and deliver them to the correct organelles^{1,2}. By contrast, many types of lipid synthesized in the endoplasmic reticulum (ER) have been suggested to be sorted to other organelles by non-vesicular mechanisms, although some lipid flux routes such as the endocytosis of plasma membrane lipids occur by vesicle-mediated mechanisms^{3–5}. Little is known, however, about the molecular machinery responsible for the interorganelle movement of lipids.

In mammalian cells, ceramide is synthesized in the ER and translocated to the Golgi compartment for conversion to sphingomyelin (SM). There are at least two pathways by which ceramide is transported from the ER to the Golgi site for the synthesis of SM: an ATP- and cytosol-dependent major pathway, and an ATP- or cytosol-independent (or less dependent) minor pathway^{6–9}. The major pathway is impaired in a Chinese hamster ovary (CHO) mutant cell line, LY-A, without any deficiency in the ER-to-Golgi trafficking of proteins^{6–8}.

In this study, we identify a factor defective in LY-A cells by functional rescue experiments and show that this factor, which we have named CERT, mediates the ATP-dependent pathway of ceramide transport from the ER to the Golgi in a non-vesicular manner.

Molecular cloning of CERT

In mammalian cells, both SM and cholesterol are preferentially distributed to the exocytosomal leaflet of the plasma membrane. Because cholesterol has a stronger affinity for SM than for glycerophospholipids, altering the concentration of coexisting SM affects the behaviour of cholesterol in artificial model membranes and biological membranes¹⁰. In fact, when exposed to the cholesterol-absorbing agent methyl- β -cyclodextrin (MCD), SM-deficient cells such as LY-A cells lose cellular cholesterol and subsequently their viability more rapidly than do wild-type cells¹¹. We exploited this phenotype for selecting revertants of LY-A cells. LY-A cells expressing an ecotropic retrovirus receptor (LY-A2 cells) were retrovirally transfected with a human complementary DNA library, and then MCD-tolerant variants were selected. We retrieved a cDNA of about

2.4 kilobases (kb) from an MCD-tolerant variant, and named the encoded protein, which has a deduced molecular mass of 68,006 Da, CERT.

Homology searching indicated that human CERT (hCERT) is identical to human GPBP Δ 26, a splicing variant of Goodpasture antigen-binding protein (GPBP). GPBP was initially identified as a protein that could bind the carboxy-terminal non-collagenous region of the human collagen α 3(IV) chain in a cell-free system^{12,13}. GPBP Δ 26, which has 26 fewer amino acid residues than GPBP, is a more common isoform and is expressed widely in various tissues, although its expression varies according to the cell type^{12,13}. Both GPBPs are hydrophilic cytoplasmic proteins^{12,13}, which conflicts with their proposed pathological role in the phosphorylation of extracellular collagen. Thus, the physiological roles of GPBPs remain unknown.

CERT fully complements LY-A cells

The transfection of LY-A2 cells with hCERT cDNA efficiently endowed the cells with MCD tolerance. A purified transfectant cell line, LY-A2/hCERT, was analysed further. Similar to wild-type CHO cells, LY-A2/hCERT cells responded to MCD and to the SM-binding cytolysin lysenin (Fig. 1a). Consistent with this, metabolic labelling experiments with radioactive serine (Fig. 1b) and choline (data not shown) indicated a recovery of the conversion of ceramide to SM in LY-A2/hCERT cells. When analysed with radioactive serine, LY-A2 cells showed a mild deficiency in the metabolic labelling of glucosylceramide (GlcCer) and ceramide (Fig. 1b). This is possibly due to a repression of the formation of sphingoid bases as a compensatory response to the reduced consumption of ceramide, although the mechanism of repression remains unknown.

To examine the conversion of ceramide to complex sphingolipids more specifically, we also carried out metabolic labelling with radioactive sphingosine, which allows the *de novo* synthesis of sphingoid bases to be bypassed in the labelling of ceramide⁷. LY-A2 cells were defective in the conversion of ceramide to SM but not to GlcCer, and this deficiency was rescued in LY-A2/hCERT cells (Fig. 1c). Chemical quantification of phospholipids also indicated the recovery of SM content in LY-A2/hCERT cells (Fig. 1d). Enzyme assays ruled out the possibility of an upregulation of SM synthase in LY-A2/hCERT cells (Fig. 1e).

Monitoring the redistribution of DMB-Cer, a fluorescent analogue of ceramide, from cytosolic reticular membranes to the

perinuclear region can be used to track ceramide movement from the ER to the Golgi^{7,14}. The DMB-Cer redistribution in LY-A2/hCERT cells was virtually identical to that in wild-type cells, whereas it was impaired in LY-A2 cells (Fig. 1f). When intracellular ATP was depleted by energy inhibitors, the ER-to-Golgi redistribution of DMB-Cer was inhibited in LY-A2/hCERT cells as it was in wild-type cells (Fig. 1f), eliminating the possibility of an upregulation of the ATP-independent minor pathway in LY-A2/hCERT cells.

Collectively, these results indicate that expression of hCERT fully complements the deficiency in the ATP-dependent pathway of ceramide transport from the ER to the Golgi in LY-A cells. Notably, expression of GPBP (referred to hereafter as CERT_L; Fig. 2a), a larger splicing variant of CERT, rescued LY-A2 cells (data not

shown), indicating that CERT_L is also functional in ceramide transport.

CERT extracts ceramide from membranes

CERT contains a phosphoinositide-binding pleckstrin homology (PH) domain in its amino-terminal region of about 120 amino acids^{12,13}, and a putative lipid-transfer-catalysing domain named START in its C-terminal region of about 230 amino acids (ref. 15 and Fig. 2a). The PH domain of this protein targets the Golgi apparatus by recognizing PtdIns4P¹⁶. The remaining middle region (MR) of CERT contains coiled-coil motifs^{12,13}, which might have a role in self-oligomerization.

To test the possibility that CERT mediates the non-vesicular

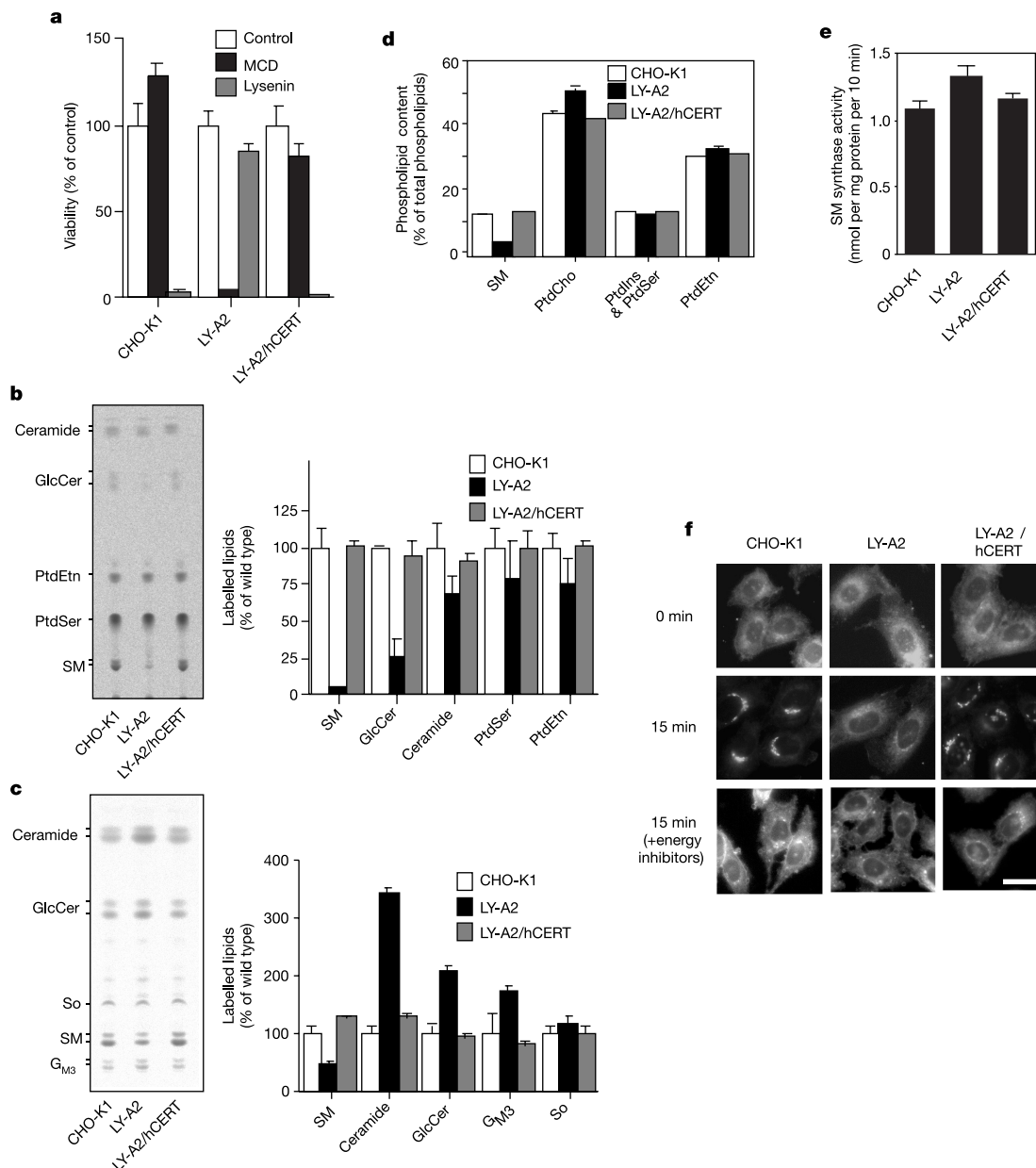


Figure 1 CERT complements the mutant phenotypes of LY-A2 cells. **a**, MCD and lysenin sensitivities of CHO cell lines. **b**, Metabolic labelling of lipids with [¹⁴C]serine. After 2 h of labelling, lipids were subjected to thin-layer chromatography (TLC) and image analysis. PtdEtn, phosphatidylethanolamine; PtdSer, phosphatidylserine. **c**, Metabolic labelling of lipids with [³H]sphingosine. After 2 h of labelling, lipids were subjected to TLC and image analysis. G_{M3}, N-acetyl neuraminyl lactosylceramide; So, sphingosine. **d**, Phospholipid

composition. PtdIns, phosphatidylinositol. **e**, SM synthase activity in the microsomal fraction. **f**, Redistribution of DMB-Cer in intact cells. Cells were labelled with 1 μ M DMB-Cer at 4 °C for 30 min, washed and chased at 33 °C for 15 min. For energy inhibition, cells were pretreated with energy inhibitors (50 mM deoxy-D-glucose and 5 mM NaN₃) at 33 °C for 15 min, chilled, labelled and chased in the presence of the energy inhibitors. Scale bar, 20 μ m.

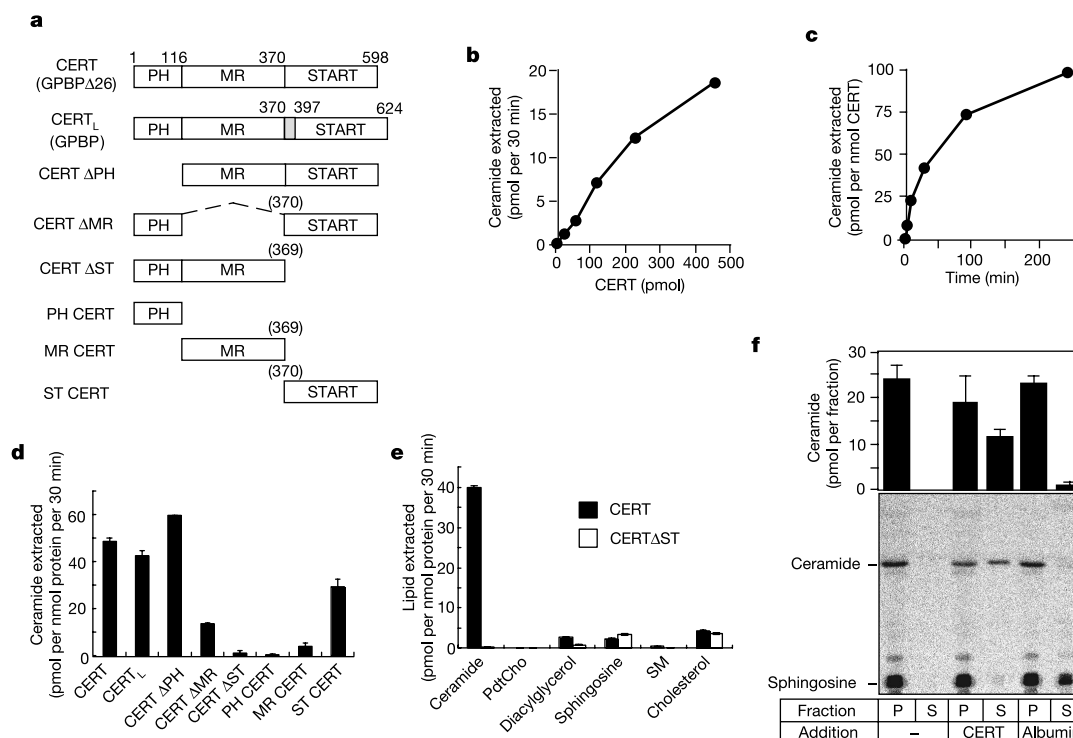


Figure 2 CERT specifically extracts ceramide from phospholipid bilayers. **a**, Structure of CERT, CERT_L and deletion mutants. In some deletion constructs the last residue (Lys 370) of the MR domain was placed in the START domain because proteins may be destabilized by the presence of a C-terminal lysine⁴⁰. **b**, CERT dependence of the extraction of C₁₆ ceramide from phospholipid vesicles. **c**, Time course of the CERT-dependent extraction of C₁₆ ceramide from phospholipid vesicles. **d**, Ceramide extraction assay with various

CERT-related constructs. **e**, Substrate specificity of lipid extraction from phospholipid vesicles by CERT. **f**, Extraction of C₁₆ ceramide from isolated ER membranes. Membranes were preincubated with palmitoyl CoA and [³H]sphingosine to form [³H]ceramide, washed, incubated with or without CERT (450 pmol) and bovine serum albumin (4.5 nmol) and centrifuged. Lipids of the pellet (P) and supernatant (S) fractions were then subjected to TLC and image analysis. Recombinant human CERT was used in **b–f**.

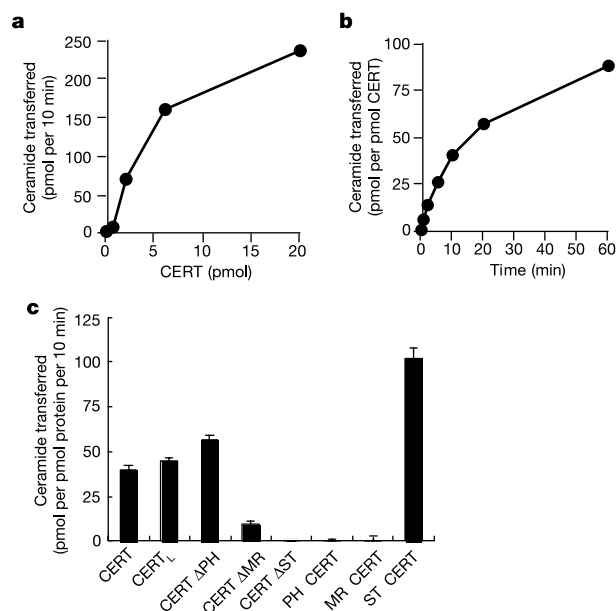


Figure 3 CERT catalyzes the intermembrane transfer of ceramide. **a**, CERT dependence of the C₁₆ ceramide transfer from donor to acceptor phospholipid vesicles. **b**, Time course of the CERT-dependent transfer of ceramide from donor to acceptor phospholipid vesicles. **c**, Ceramide transfer assay with various CERT-related constructs. Recombinant human CERT was used in all experiments shown.

transport of ceramide, we examined whether CERT could extract ceramide from phospholipid bilayers in a cell-free system. In this assay, we used natural long-chain ceramide as the substrate. Purified recombinant CERT efficiently extracted ceramide from phospholipid vesicles in a dose- and time-dependent manner (Fig. 2b, c). Incubation with 450 pmol of CERT for 30 min resulted in the extraction of about 20% of the 225 pmol of ceramide initially loaded into the vesicles (Fig. 2c). Deletion of the START domain abrogated the activity completely, whereas mutants in which the PH or the MR domain was deleted retained the activity (Fig. 2d). In addition, the START domain alone showed substantial activity, whereas neither the PH domain alone nor the MR domain alone did (Fig. 2d). The activity of CERT_L was almost the same as that of CERT (Fig. 2d).

With regard to other proteins containing START domains, StAR and MLN64 bind cholesterol^{17,18}, and the phosphatidylcholine (PtdCho)-transfer protein binds PtdCho¹⁹. Using the START deletion as a negative control, we found that CERT could not extract sphingosine, SM, PtdCho or cholesterol from phospholipid vesicles (Fig. 2e). The activity of CERT towards extracting diacylglycerol, a glycerolipid that structurally resembles ceramide, was about 5% of its activity towards extracting ceramide (Fig. 2e).

We further tested whether CERT can indeed extract ceramide from the ER. When ER membranes with long-chain [³H]ceramide endogenously formed from palmitoyl CoA and [³H]sphingosine were incubated with recombinant CERT for 30 min, about 40% of [³H]ceramide was extracted from the ER (Fig. 2f). CERT did not extract [³H]sphingosine even though the ER contained an excess

amount of the precursor [^3H]sphingosine (Fig. 2f). By contrast, albumin, a nonspecific lipid adsorbent, efficiently extracted [^3H]sphingosine, but not [^3H]ceramide (Fig. 2f). These results indicate that CERT interacts with the ER membranes and specifically extracts ceramide.

CERT mediates the intermembrane transfer of ceramide

The spontaneous transfer of long-chain C_{16} ceramide between phospholipid vesicles is very slow, taking of the order of days²⁰. When we assayed the transfer of C_{16} ceramide between phospholipid vesicles in a cell-free system, we found that CERT greatly facilitated the intermembrane transfer of ceramide (Fig. 3a, b). One picomole of CERT catalysed the transfer of about 40 pmol of ceramide from donor to acceptor vesicles in 10 min, whereas no transfer of ceramide occurred in the absence of CERT (Fig. 3a, b). An analysis of deletion mutant constructs showed that the START domain was necessary and sufficient for catalysing the transfer of ceramide between phospholipid vesicles (Fig. 3c), consistent with the results on the ceramide extraction activity (Fig. 2d).

When donor vesicles doubly labelled with [^{14}C]ceramide and [^3H]PtdCho were used, a CERT-dependent transfer of [^{14}C]ceramide, but not [^3H]PtdCho, was observed (data not shown), confirming the specific transfer of ceramide. Thus, CERT can catalyse the specific extraction of ceramide from donor membranes and, dependent on its START domain, can transfer the extracted ceramide to acceptor membranes *in vitro*.

The LY-A cell line expresses mutated CERT

We isolated cDNAs encoding the hamster homologue of CERT from both wild-type CHO-K1 cells and the mutant cell line LY-A. The CERT derived from the CHO-K1 cell line (cCERT) was 98% identical to human and mouse CERT at the amino acid level. The LY-A-derived CERT cDNA was identical to cCERT except for a change at codon 67 from GGA (glycine; G) to GAA (glutamic acid;

E). When epitope-tagged cCERT (cCERT-Flag) and cCERT(G67E) (cCERT(G67E)-Flag) were introduced into LY-A2 cells with about 60% transfection efficiency, both constructs were equally expressed (Fig. 4c). However, cCERT-Flag, but not cCERT(G67E)-Flag, complemented the MCD hypersensitivity of the cells (Fig. 4b), indicating that the G67E mutation is the causative mutation of the phenotype of LY-A cells.

Gly 67 in CERT is located in the PH domain. This glycine residue is conserved not only among CERT or GPBPΔ26 orthologues of various multicellular organisms (Fig. 4a), but also among a PtdIns4P-recognizing subfamily of PH domains^{16,21}. A previous study has shown that the PH domain of human GPBP binds to both phosphatidylinositol-4 monophosphate (PtdIns4P) and phosphatidylinositol-4,5-diphosphate (PtdIns(4,5)P) when the lipids are on artificial phospholipid vesicles, but it probably only binds to PtdIns4P *in vivo* because its distribution is restricted to the Golgi by an unknown second determinant¹⁶. In addition, the PH domain is sufficient to confer Golgi targeting in mammalian cells¹⁶.

We examined whether the G67E mutation affects the phosphoinositide recognition activity of hCERT and cCERT by using phospholipid-spotted strips. The PH domain of both wild-type cCERT and wild-type hCERT specifically bound to PtdIns4P among various phosphoinositides; however, the PH domain of the mutant cCERT(G67E) did not recognize PtdIns4P (Fig. 4d). Regardless of the G67E mutation, full-size CERT proteins weakly bound to various phosphoinositides in addition to PtdIns4P. Binding assays with various deletion constructs indicated that this weak interaction was attributable to the MR and START domains (Fig. 4d).

We examined the intracellular distribution of cCERT and cCERT(G67E) by monitoring the proteins fused to green fluorescent protein (GFP). When expressed in CHO-K1 cells, cCERT-GFP was distributed throughout the cytosol but enriched in a perinuclear region, where GS28, a Golgi marker, was localized (Fig. 5).

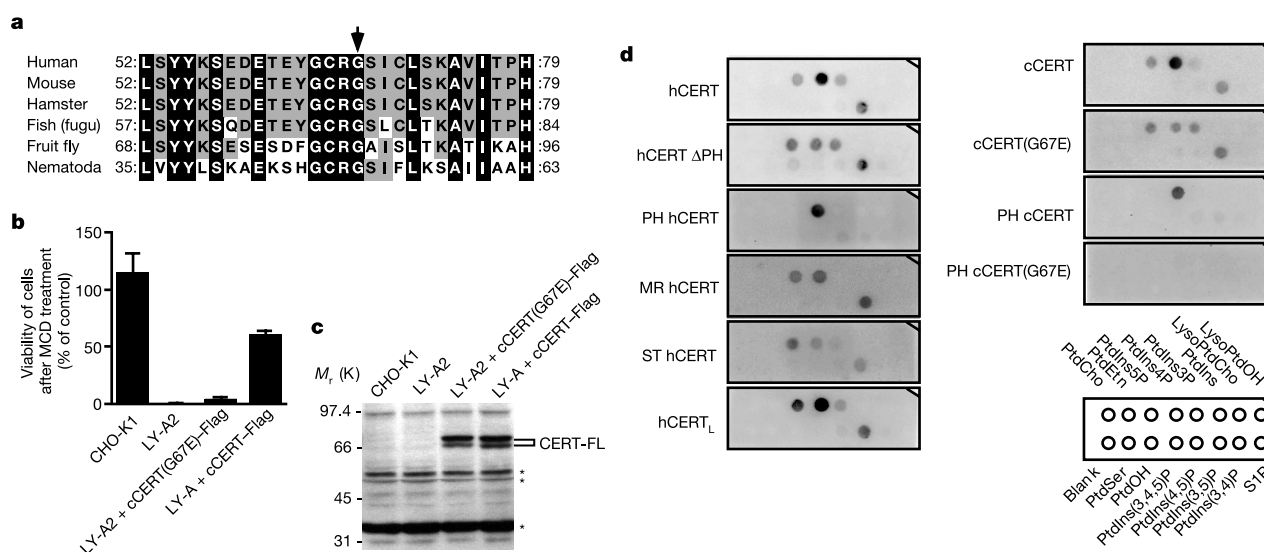


Figure 4 CERT from the mutant LY-A cell line contains a mutation. **a**, Alignment of CERT orthologues. Arrow indicates Gly 67 of mammalian CERT. The GenBank accession number of the fly *Drosophila melanogaster* orthologue is AF145644. The gene name of the nematode *Caenorhabditis elegans* orthologue is F25H2.6. The gene model number of the *Fugu rubripes* orthologue is SINFRUP0000082192. **b**, MCD sensitivity of cells. Untransfected CHO-K1 and LY-A2 cells, and LY-A2 cells transfected with cCERT-Flag or the cCERT(G67E)-Flag subcloned on a retrovirus vector were assayed. **c**, Expression of the epitope-tagged CERT constructs. Cell lysates were analysed by western blotting with M2 monoclonal antibodies to Flag. Asterisks represent unknown crossreacting proteins,

which can serve as loading controls. The doublet appearance of CERT in western blotting is due to different phosphorylation states¹³. **d**, Binding of various CERT recombinants to phospholipids. A protein-lipid overlay assay was carried out with lipid-spotted strips (100 pmol of lipid per spot). PtdOH, phosphatidic acid; PtdIns5P, phosphatidylinositol-5-monophosphate; PtdIns3P, phosphatidylinositol-3-monophosphate; PtdIns(3,4)P, phosphatidylinositol-3,4-diphosphate; PtdIns(4,5)P, phosphatidylinositol-4,5-diphosphate; PtdIns(3,4,5)P, phosphatidylinositol-3,4,5-triphosphate; S1P, sphingosine-1-phosphate.

This Golgi association was impaired by the G67E mutation (Fig. 5). The distribution patterns of the GFP fusion constructs were not due to the nature of the GFP moiety, because they were distinct from the nuclear distribution pattern of unfused GFP (Fig. 5). No degradation products of these GFP-fusion constructs were detected on

western blotting with anti-GFP monoclonal antibody (data not shown), confirming that the green fluorescence seen in Fig. 5 reflects the distribution of the full-size fusion proteins. These results indicate that CERT, dependent on its PtdIns4P-recognizing PH domain, can associate with the Golgi apparatus.

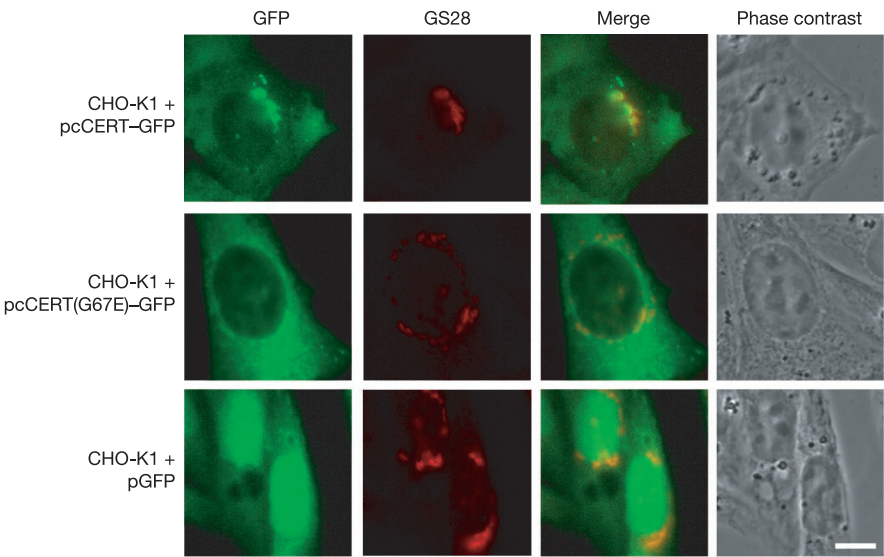


Figure 5 Intracellular distribution of GFP fusions of CERT. After transfection with expression plasmids encoding GFP alone, cCERT–GFP or cCERT(G67E)–GFP, CHO-K1 cells were processed for indirect immunostaining of GS28. Scale bar, 10 μ m.

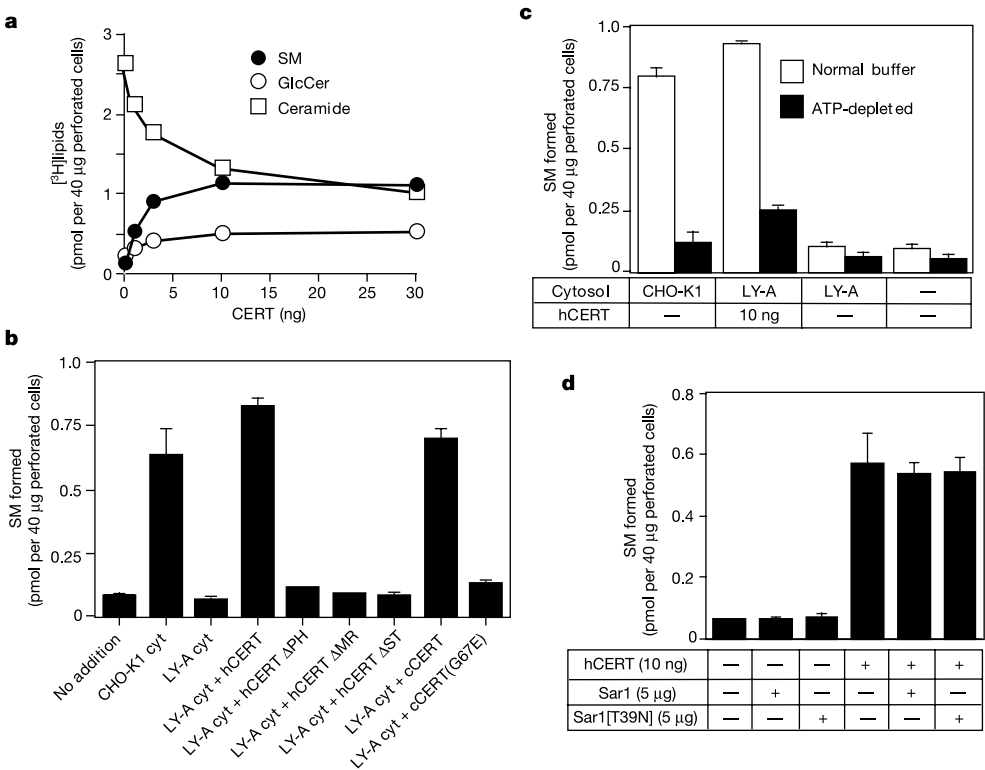


Figure 6 CERT-dependent trafficking of ceramide in semi-intact cells. **a**, CERT dependence of ceramide transport in semi-intact cells. Radioactive ceramide was formed in perforated LY-A cells. The prelabelled semi-intact cells (40 μ g) were then incubated in a chase mixture containing the LY-A cytosol (100 μ g) and various doses of recombinant hCERT at 37 $^{\circ}$ C for 30 min. **b**, Ceramide transport assay with various CERT constructs. The prelabelled semi-intact cells were chased with or without the cytosol (100 μ g) of the indicated cell lines and various CERT constructs (10 ng). cyt, cytosol. **c**, Effects of ATP

depletion on ceramide transport. Prelabelled semi-intact cells were chased with or without the indicated cytosol (100 μ g) and recombinant hCERT (10 ng). For ATP depletion, apyrase (1 unit) was added in the chase mixture in place of the ATP regeneration system. **d**, Effects of dominant-negative Sar1 on ceramide transport. Prelabelled semi-intact cells were chased with the LY-A cell cytosol (100 μ g) in the presence or absence of hCERT (10 ng) and either Sar1 (5 μ g) or Sar1[T39N] (5 μ g).

Reconstitution in semi-intact cells

To examine further the CERT-dependent transport of ceramide, we analysed ceramide movement from the ER to the Golgi by using a semi-intact cell system, which reproduces various aspects of the ceramide-trafficking events that occur in intact cells⁸. In agreement with our previous study⁸, the defect in ceramide transport from the ER to the site of SM synthesis in LY-A cells was rescued by replacing the mutant cytosol with wild-type cytosol (Fig. 6b). The defect was also rescued by adding purified recombinant CERT to the mutant cytosol. Addition of 10 ng (~0.14 pmol) of CERT to 100 µg of mutant cytosol enhanced the ceramide-to-SM conversion by about tenfold, recovering the activity to that of the wild type (Fig. 6a, b).

This result eliminated the possibility that CERT affects ceramide synthesis, thereby inhibiting SM synthesis, because the *in vitro* transport assay begins after ceramide is synthesized. In contrast to the synthesis of SM, conversion of ceramide to GlcCer in semi-intact cells can occur with either the wild-type or the LY-A cytosol, and even without cytosol⁸. The maximal enhancement of ceramide-to-GlcCer conversion on addition of CERT was only about twofold (Fig. 6a), suggesting that the contribution of CERT to the conversion of ceramide to GlcCer is minor.

Notably, neither the PH deletion mutant nor the G67E mutant rescued the conversion of ceramide to SM in semi-intact cells (Fig. 6b), although these mutants were similar to wild-type CERT in terms of their ceramide extraction and transfer activities, as assayed in cell-free systems using artificial phospholipid vesicles (Figs 2d and 3c; data not shown for the G67E mutant). Presumably, the PH domain of CERT is crucial for selective targeting to the Golgi when various other organelles coexist, as they do in intact cells. Deletion of the MR domain did not rescue the conversion of ceramide to SM. This may be relevant to a possible ER-targeting function of the MR domain (see below).

Depletion of ATP strongly inhibited the CERT-dependent conversion of ceramide to SM in the semi-intact cell system (Fig. 6c), which is consistent with our above result that ATP depletion inhibited the CERT-dependent movement of ceramide from the ER to the Golgi in intact cells (Fig. 1f). Sar1 protein is an essential factor for the vesicle-mediated protein transport from the ER to the Golgi apparatus²², and its dominant-negative form, Sar1[T39N], can inhibit vesicle trafficking from the ER to the Golgi²³. Neither mammalian Sar1 nor Sar1[T39N] affected the conversion of ceramide to SM (Fig. 6d), providing more evidence that CERT-dependent transport of ceramide is non-vesicular.

Discussion

We have presented genetic and biochemical evidence that CERT is central to ceramide-selective transport during the synthesis of SM. To our knowledge, our study is the first example of the molecular identification of a specific lipid-sorting factor that operates in the synthesis of membrane phospholipids. For the transport of ceramide from the ER to the Golgi, CERT has two suitable domains: the START domain, which specifically recognizes ceramide and mediates the intermembrane transfer of ceramide; and the PH domain, which can target the Golgi apparatus where several types of PtdIns-4-OH kinase are localized^{24,25}. In addition, the MR domain of CERT contains a motif for targeting the ER²⁶.

It is possible that CERT has a regulatory role in ceramide transport and binds ceramide only to monitor its intracellular concentrations; however, CERT can catalyse the intermembrane transfer of ceramide in a cell-free system (Fig. 3) and the G67E mutation in the PH domain inhibits the conversion of ceramide to SM in both intact (Fig. 1c) and semi-intact (Fig. 6b) cells. These results are best explained by the notion that CERT mediates the transport of ceramide from the ER to the Golgi. Thus, we propose the model that CERT extracts newly synthesized ceramide from the ER and carries the ceramide molecule, in a non-vesicular manner, to

the Golgi apparatus. As a specific subdomain of the ER is suggested to be closely associated with Golgi stacks²⁷, CERT-mediated transfer might occur efficiently at the ER–Golgi contact site. The ATP dependence of ceramide transport in cells could be relevant to the metabolism of PtdIns4P at the Golgi apparatus or the phosphorylation of CERT. Further studies are needed to elucidate how the recycling of CERT between different organelles is regulated.

In contrast to the synthesis of SM, the ATP- and cytosol-dependent pathway of ceramide transport makes a minor contribution to the *de novo* synthesis of GlcCer in CHO cells^{7,8} and in rat C6 glioma cells²⁸. Consistent with our previous studies^{7,8}, LY-A2 cells showed no obvious deficiency in the conversion of ceramide to GlcCer (Fig. 1c), supporting the idea that the access of ceramide to the site of GlcCer synthesis is almost normal in the mutant cells. These results suggest that ceramide is accessible to the main site of *de novo* synthesis of GlcCer in a CERT-independent manner, at least in CHO cells. This might be simply due to a partial distribution of GlcCer synthase to the ER²⁹, as discussed previously⁸. Alternatively, a specific pool of ceramide molecules might be selectively translocated to the site of GlcCer synthesis by an unknown mechanism. In support of the latter possibility, overexpression of the *uog1* gene in human embryonic kidney 293T cells results in an upregulation of *de novo* synthesis of C₁₈ ceramide, but not of C₁₆ ceramide, and the C₁₈ ceramide synthesized is preferentially used for the synthesis of GlcCer³⁰.

Invertebrate animals have CERT orthologues (refs 12, 15, and Fig. 4a), although it remains to be determined whether these orthologues can recognize ceramide. We failed to find a clear relative of CERT in yeast with the publicly available tools for homology searches, even though ceramide transport from the ER to the Golgi must occur in this organism³¹. This may indicate that additional mechanisms exist in other species, or that this role could be taken over by more distantly related proteins. Although START domains are probably absent in yeast³², it should be noted that yeast Osh1p/Swh1p and Osh2p, which are members of a family containing a lipid-binding domain related to mammalian oxysterol-binding protein, have both a PtdIns4P-recognizing PH domain and an ER-targeting motif that are similar to those of CERT^{16,26}.

Ceramide, which can be produced in both anabolic and catabolic pathways, is a modulator of various cellular events^{33,34}. When ceramide exerts its bio-modulator activities at organelles that differ from the ceramide-producing organelles, interorganelle transport of ceramide must occur. CERT might be also involved in the interorganelle transfer of ceramide for ceramide-mediated signalling. □

Methods

Gene cloning

The LY-A2 cell line was established by the stable transfection of LY-A cells with mCAT-1, which encodes the murine ecotropic retrovirus receptor³⁵. We prepared retroviral particles with a human HeLa retroviral library kit (Clontech) and Plat-E packaging cells³⁶. After infection for 6 h, LY-A2 cells were cultured at 37 °C for 16 h, reseeded, and cultured at 33 °C overnight. After being washed with serum-free F-12 medium, the cells were incubated in 10 mM MCD in F-12 at 37 °C for 1 h, washed with PBS, replenished with culture medium, and cultured at 33 °C for 7–10 d. Surviving cells were reseeded and subjected again to MCD treatment. After a total of four cycles of MCD treatment, an MCD-tolerant variant of LY-A2 was purified by limiting dilution.

Using genomic polymerase chain reaction (PCR) with an LA-PCR kit (Takara) and primers 1 and 2 (PCR primers are listed in Supplementary Information), we amplified, cloned and sequenced the cDNA integrated into the genome of the variant. The 1.8-kb open-reading frame (ORF) of human CERT amplified by PCR (template, human HeLa retroviral library; primers, 3 and 5) was cloned into pBlueScript (Stratagene) to obtain pBS/hCERT, and subcloned into a pLIB retroviral vector (Clontech). The human CERT_L/GPBP ORF was constructed as described in Supplementary Information. CERT cDNAs from CHO-K1 and LY-A cells were obtained by PCR with reverse transcription (Supplementary Information). Procedures for producing constructs encoding various deletion mutants, GFP-fused CERTs and epitope-tagged CERTs are also described in Supplementary Information.

Purification of recombinant proteins

CERT, CERT_L and deletion mutant constructs (see Fig. 2a) were subcloned into pET-28a(+) (Novagen) to express the proteins with a His₆ tag at the N terminus. These

plasmids were transfected into BL21 (DE3) *Escherichia coli* cells. We prepared lysate from the transfected cells as described²¹, except that we omitted divalent cation chelators from the lysis buffer and reduced the concentration of 2-mercaptoethanol to 2.5 mM. His₆-tagged proteins were purified from the lysate using a Talon Co²⁺ affinity column (Clontech). His₆-tagged recombinants of mammalian Sar1 and Sar1[T39N] expressed in M15/pREP4 *E. coli* cells were similarly purified.

Assay of ceramide-extracting activity

Large liposomes consisting of egg-yolk PtdCho, phosphatidylethanolamine and [*palmitoyl*-1-¹⁴C] *N*-palmitoyl-D-erythro-sphingosine ([¹⁴C]ceramide) (800:200:3, mol:mol:mol) were prepared by mild sonication as described³⁷. In the standard assay, lipid vesicles (50 µg) and purified recombinant hCERT (450 pmol) were incubated in 100 µl of 50 mM HEPES-NaOH (pH 7.4) containing 100 mM NaCl and 0.5 mM EDTA at 37 °C for 30 min. After centrifugation (50,000 g, 30 min), the radioactivity of the supernatant was determined as a measure of the ceramide extracted from phospholipid bilayers. Background radioactivity from CERT-free controls was subtracted from the radioactivity of each sample (see Supplementary Information for additional information). The method of extraction of ceramide from isolated ER membranes is also described in Supplementary Information.

Other methods

Lipid analyses (metabolic labelling, phospholipid composition, intracellular DMB-Cer redistribution and SM synthase activity) were done as described^{6,7,9}. Sensitivities of cells to MCD and lysein were determined as described⁶. The transfer of C₁₆ ceramide between artificial phospholipid vesicles was assayed by a modification of a previously reported method³⁸, as described in Supplementary Information.

We assayed the binding of purified CERT recombinants to phosphoinositides using PIP-Strips (Echelon Biosciences) as described in Supplementary Information. GFP fusion constructs were transiently expressed in CHO cells by lipofection. Immunocytochemical analysis was carried out with monoclonal antibodies against GS28 (StressGen) and Alexa-594-conjugated secondary antibodies (Molecular Probe) as described³⁹. The reconstitution of ceramide transport from the ER to the Golgi in semi-intact cells was done as described⁸.

Data analysis

Data are presented as the mean ± s.e.m. or are images typical of at least three experiments.

Received 18 September; accepted 30 October 2003; doi:10.1038/nature02188.

- Chen, Y. A. & Scheller, R. H. SNARE-mediated membrane fusion. *Nature Rev. Mol. Cell Biol.* **2**, 98–106 (2001).
- Zerial, M. & McBride, H. Rab proteins as membrane organizers. *Nature Rev. Mol. Cell Biol.* **2**, 107–117 (2001).
- Voelker, D. R. In *Biochemistry of Lipids, Lipoproteins and Membranes* 4th edn (eds Vance, D. E. & Vance, J.) 449–482 (Elsevier, Amsterdam, 2002).
- van Meer, G. & Holthuis, J. C. M. Sphingolipid transport in eukaryotic cells. *Biochim. Biophys. Acta* **1486**, 145–170 (2000).
- Maxfield, F. R. & Wüstner, D. Intracellular cholesterol transport. *J. Clin. Invest.* **110**, 891–898 (2002).
- Hanada, K. *et al.* Mammalian cell mutants resistant to a sphingomyelin-directed cytolysin. Genetic and biochemical evidence for complex formation of the LCB1 protein with the LCB2 protein for serine palmitoyltransferase. *J. Biol. Chem.* **273**, 33787–33794 (1998).
- Fukasawa, M., Nishijima, M. & Hanada, K. Genetic evidence for ATP-dependent endoplasmic reticulum-to-Golgi apparatus trafficking of ceramide for sphingomyelin synthesis in Chinese hamster ovary cells. *J. Cell Biol.* **144**, 673–685 (1999).
- Funakoshi, T., Yasuda, S., Fukasawa, M., Nishijima, M. & Hanada, K. Reconstitution of ATP- and cytosol-dependent transport of *de novo* synthesized ceramide to the site of sphingomyelin synthesis in semi-intact cells. *J. Biol. Chem.* **275**, 29938–29945 (2000).
- Yasuda, S. *et al.* A novel inhibitor of ceramide trafficking from the endoplasmic reticulum to the site of sphingomyelin synthesis. *J. Biol. Chem.* **276**, 43994–44002 (2001).
- Ohvo-Rekilä, H., Ramstedt, B., Leppimäki, P. & Slotte, J. P. Cholesterol interactions with phospholipids in membranes. *Prog. Lipid Res.* **41**, 66–97 (2002).
- Fukasawa, M., Nishijima, M., Itabe, H., Takano, T. & Hanada, K. Reduction of sphingomyelin level without accumulation of ceramide in Chinese hamster ovary cells affects detergent-resistant membrane domains and enhances cellular cholesterol efflux to methyl-β-cyclodextrin. *J. Biol. Chem.* **275**, 34028–34034 (2000).
- Raya, A., Revert, F., Navarro, S. & Saus, J. Characterization of a novel type of serine/threonine kinase that specifically phosphorylates the human goodpasture antigen. *J. Biol. Chem.* **274**, 12642–12649 (1999).
- Raya, A. *et al.* Goodpasture antigen-binding protein, the kinase that phosphorylates the goodpasture antigen, is an alternatively spliced variant implicated in autoimmune pathogenesis. *J. Biol. Chem.* **275**, 40392–40399 (2000).
- Pagano, R. E., Martin, O. C., Kang, H. C. & Haugland, R. P. A novel fluorescent ceramide analogue for studying membrane traffic in animal cells: Accumulation at the Golgi apparatus results in altered spectral properties of the sphingolipid precursor. *J. Cell Biol.* **113**, 1267–1279 (1991).
- Ponting, C. P. & Aravind, L. START: a lipid-binding domain in StAR, HD-ZIP and signalling proteins. *Trends Biochem. Sci.* **24**, 130–132 (1999).
- Levine, T. P. & Munro, S. Targeting of Golgi-specific pleckstrin homology domains involves both PtdIns 4-kinase-dependent and -independent components. *Curr. Biol.* **12**, 695–704 (2002).
- Kallen, C. B. *et al.* Steroidogenic acute regulatory protein (StAR) is a sterol transfer protein. *J. Biol. Chem.* **273**, 26285–26288 (1998).
- Tsujiushita, Y. & Hurley, J. H. Structure and lipid transport mechanism of a StAR-related domain. *Nature Struct. Biol.* **7**, 408–414 (2000).
- Akeroyd, R., Moonen, P., Westerman, J., Puyk, W. C. & Wirtz, K. W. The complete primary structure of the phosphatidylcholine-transfer protein from bovine liver. Isolation and characterization of the cyanogen bromide peptides. *Eur. J. Biochem.* **114**, 385–391 (1981).
- Simon, C. G. Jr, Holloway, P. W. & Gear, A. R. Exchange of C₁₆-ceramide between phospholipid vesicles. *Biochemistry* **38**, 14676–14682 (1999).
- Dowler, S. *et al.* Identification of pleckstrin-homology-domain-containing proteins with novel phosphoinositide-binding specificities. *Biochem. J.* **351**, 19–31 (2000).
- Nakano, A. & Muramatsu, M. A novel GTP-binding protein, Sar1p, is involved in transport from the endoplasmic reticulum to the Golgi apparatus. *J. Cell Biol.* **109**, 2677–2691 (1989).
- Kuge, O. *et al.* Sar1 promotes vesicle budding from the endoplasmic reticulum but not Golgi compartments. *J. Cell Biol.* **125**, 51–65 (1994).
- De Matteis, M., Godi, A. & Corda, D. Phosphoinositides and the Golgi complex. *Curr. Opin. Cell Biol.* **14**, 434–447 (2002).
- Wang, Y. J. *et al.* Phosphatidylinositol 4 phosphate regulates targeting of clathrin adaptor AP-1 complexes to the Golgi. *Cell* **114**, 299–310 (2003).
- Loewen, C. J., Roy, A. & Levine, T. P. A conserved ER targeting motif in three families of lipid binding proteins and in Opi1p binds VAP. *EMBO J.* **22**, 2025–2035 (2003).
- Ladinsky, M. S., Mastronarde, D. N., McIntosh, J. R., Howell, K. E. & Staehelin, L. A. Golgi structure in three dimensions: functional insights from the normal rat kidney cell. *J. Cell Biol.* **144**, 1135–1149 (1999).
- Viani, P. *et al.* Ceramide in nitric oxide inhibition of glioma cell growth. Evidence for the involvement of ceramide traffic. *J. Biol. Chem.* **278**, 9592–9601 (2003).
- Ardail, D. *et al.* The mitochondria-associated endoplasmic-reticulum subcompartment (MAM fraction) of rat liver contains highly active sphingolipid-specific glycosyltransferases. *Biochem. J.* **371**, 1013–1019 (2003).
- Venkataraman, K. *et al.* Upstream of growth and differentiation factor 1 (*uog1*), a mammalian homolog of the yeast longevity assurance gene 1 (*LAG1*), regulates *N*-stearoyl-sphinganine (C₁₈-(dihydro)ceramide) synthesis in a fumonisin B1-independent manner in mammalian cells. *J. Biol. Chem.* **277**, 35642–35649 (2002).
- Funato, K. & Riezman, H. Vesicular and nonvesicular transport of ceramide from ER to the Golgi apparatus in yeast. *J. Cell Biol.* **155**, 949–959 (2001).
- Soccio, R. E. & Breslow, J. L. StAR-related lipid transfer (START) proteins: mediators of intracellular lipid metabolism. *J. Biol. Chem.* **278**, 22183–22186 (2003).
- Mathias, S., Pena, L. A. & Kolesnick, R. N. Signal transduction of stress via ceramide. *Biochem. J.* **335**, 465–480 (1998).
- Hannun, Y. A. & Luberto, C. Ceramide in the eukaryotic stress response. *Trends Cell Biol.* **10**, 73–80 (2000).
- Albritton, L. M., Tseng, L., Scadden, D. & Cunningham, J. M. A putative murine ecotropic retrovirus receptor gene encodes a multiple membrane-spanning protein and confers susceptibility to virus infection. *Cell* **57**, 659–666 (1989).
- Morita, S., Kojima, T. & Kitamura, T. Plat-E: an efficient and stable system for transient packaging of retroviruses. *Gene Ther.* **7**, 1063–1066 (2000).
- Patki, V. *et al.* Identification of an early endosomal protein regulated by phosphatidylinositol 3-kinase. *Proc. Natl Acad. Sci. USA* **94**, 7326–7330 (1997).
- Kasper, A. M. & Helmkamp, G. M. Jr Intermembrane phospholipid fluxes catalyzed by bovine brain phospholipid exchange protein. *Biochim. Biophys. Acta* **664**, 22–32 (1981).
- Yasuda, S., Nishijima, M. & Hanada, K. Localization, topology, and function of the LCB1 subunit of serine palmitoyltransferase in mammalian cells. *J. Biol. Chem.* **278**, 4176–4183 (2003).
- Banerjee, A., Gregori, L., Xu, Y. & Chau, V. The bacterially expressed yeast CDC34 gene product can undergo autoubiquitination to form a multiubiquitin chain-linked protein. *J. Biol. Chem.* **268**, 5668–5675 (1993).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank J. Cunningham for the mCAT-1 cDNA; T. Kitamura for the Plat-E cells; O. Kuge for the 5'-RACE- and 3'-RACE-ready CHO cDNAs and the *E. coli* expression plasmids for CHO Sar1a and Sar1a[T39N]; and Y. Sekizawa for lysein. This work was supported by the Ministry of Education, Culture, Sports, Science and Technology (MEXT).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to K.H. (hanak@nih.go.jp).

A gravitationally lensed quasar with quadruple images separated by 14.62 arcseconds

Naohisa Inada¹, Masamune Oguri¹, Bartosz Pindor², Joseph F. Hennawi², Kuenley Chiu³, Wei Zheng³, Shin-ichi Ichikawa⁴, Michael D. Gregg^{5,6}, Robert H. Becker^{5,6}, Yasushi Suto¹, Michael A. Strauss², Edwin L. Turner², Charles R. Keeton⁷, James Annis⁸, Francisco J. Castander⁹, Daniel J. Eisenstein¹⁰, Joshua A. Frieman^{7,8}, Masataka Fukugita¹¹, James E. Gunn², David E. Johnston⁷, Stephen M. Kent⁸, Robert C. Nichol¹², Gordon T. Richards², Hans-Walter Rix¹³, Erin Scott Sheldon⁷, Neta A. Bahcall², J. Brinkmann¹⁴, Željko Ivezić², Don Q. Lamb⁷, Timothy A. McKay¹⁵, Donald P. Schneider¹⁶ & Donald G. York^{7,17}

¹Department of Physics, School of Science, The University of Tokyo, 113-0033, Japan

²Princeton University Observatory, Peyton Hall, Princeton, New Jersey 08544, USA

³Department of Physics and Astronomy, Johns Hopkins University, 3701 San Martin Drive, Baltimore, Maryland 21218, USA

⁴National Astronomical Observatory, 2-21-1 Osawa, Mitaka, Tokyo 181-8588, Japan

⁵Department of Physics, University of California at Davis, 1 Shields Avenue, Davis, California 95616, USA

⁶Institute of Geophysics and Planetary Physics, Lawrence Livermore National Laboratory, L-413, 7000 East Avenue, Livermore, California 94550, USA

⁷Department of Astronomy and Astrophysics, University of Chicago, 5640 South Ellis Avenue, Chicago, Illinois 60637, USA

⁸Fermi National Accelerator Laboratory, PO Box 500, Batavia, Illinois 60510, USA

⁹Institut d'Estudis Espacials de Catalunya/CSIC, Gran Capita 2-4, 08034 Barcelona, Spain

¹⁰Steward Observatory, University of Arizona, 933 North Cherry Avenue, Tucson, Arizona 85721, USA

¹¹Institute for Cosmic Ray Research, The University of Tokyo, 5-1-5 Kashiwa, Chiba 277-8582, Japan

¹²Department of Physics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA

¹³Max-Planck Institute for Astronomy, Königstuhl 17, D-69117 Heidelberg, Germany

¹⁴Apache Point Observatory, PO Box 59, Sunspot, New Mexico 88349, USA

¹⁵Department of Physics, University of Michigan, 500 East University Avenue, Ann Arbor, Michigan 48109, USA

¹⁶Department of Astronomy and Astrophysics, Pennsylvania State University, 525 Davey Laboratory, University Park, Pennsylvania 16802, USA

¹⁷Enrico Fermi Institute, University of Chicago, 5640 South Ellis Avenue, Chicago, Illinois 60637, USA

Gravitational lensing is a powerful tool for the study of the distribution of dark matter in the Universe. The cold-dark-matter model of the formation of large-scale structures (that is, clusters of galaxies and even larger assemblies) predicts^{1–6} the existence of quasars gravitationally lensed by concentrations of dark matter⁷ so massive that the quasar images would be split by over 7 arcsec. Numerous searches^{8–11} for large-separation lensed quasars have, however, been unsuccessful. All of the roughly 70 lensed quasars known¹², including the first lensed quasar discovered¹³, have smaller separations that can be explained in terms of galaxy-scale concentrations of baryonic matter. Although gravitationally lensed galaxies¹⁴ with large separations are known, quasars are more useful cosmological probes because of the simplicity of the resulting lens systems. Here we report the discovery of a lensed quasar, SDSS J1004+4112, which has a maximum separation between the components of 14.62 arcsec. Such a large separation means that the lensing object must be dominated by dark matter. Our results are fully consistent with theoretical expectations^{3–5} based on the cold-dark-matter model.

For bright quasars (i-band magnitude $i < 19$), the probability of gravitational lensing¹⁵ is only about 0.1%; the majority of these lenses have small separations, due to a single massive galaxy. The fraction of large-separation lensed quasars is predicted to be 0.01% or less^{3–5}; thus it is not surprising that none have been found to date^{8–11}. In order to find such objects, we need samples of tens of thousands of quasars, such as those generated by the Sloan Digital Sky Survey^{16,17} (SDSS). The SDSS is conducting both a photometric survey^{18–22} using five broad optical bands²² (u, g, r, i and z) and a spectroscopic survey²³ of 10,000 square degrees of the sky centred approximately on the North Galactic Pole, using a dedicated wide-field 2.5-m telescope at the Apache Point Observatory.

We searched for large-separation lensed quasars in a sample of ~29,500 spectroscopically confirmed SDSS quasars²⁴ at redshifts z of 0.6–2.3, a sample roughly three times larger than those used in previous searches. Even with this large sample, the expected number of large-separation lensed quasars is of the order of unity. In the field around each quasar in the sample, we searched for stellar objects with colours differing by less than 0.1 from those of the quasar, with separations between 7.0" and 60.0" and with flux greater than one-tenth that of the quasar. SDSS J1004+4112 was identified as a 'quadruple' large-separation lensed quasar candidate using these criteria. Only one of the four components (component B, see below) has an SDSS spectrum (the SDSS hardware²³ does not allow pairs of objects separated by less than 55" to be observed on a single plate), and therefore, we obtained spectra of all four components using the

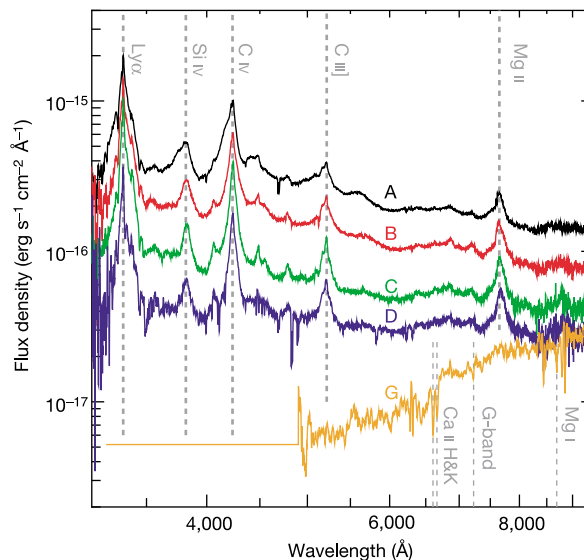


Figure 1 The Keck spectra of the four quasar components A–D, and the brightest galaxy G in the lensing cluster. See Fig. 2 for these identifications (A–D, and G). The data were taken using the Low-Resolution Imaging Spectrometer²⁹ of the Keck I telescope. The exposure times were 900 s for each component. The dispersion is 1.09 Å pixel^{−1}. The data were reduced in a standard method using IRAF (IRAF is the image reduction and analysis facility, distributed by the National Optical Astronomy Observatories). The black solid line, the red solid line, the green solid line and the blue solid line represent the spectra of components A, B, C and D, respectively. The vertical grey dashed lines with labels at the top of the figure (3,323.6 Å, 3,818.7 Å, 4,235.1 Å, 5,218.5 Å and 7,651.8 Å) represent the positions of emission lines of the respective ions redshifted to $z = 1.734$ of Ly α (1,215.67 Å), Si IV (1,396.76 Å), C IV (1,549.06 Å), C III (1,908.73 Å) and Mg II (2,798.75 Å), respectively. All emission lines are clearly at the same redshift. The orange solid line represents the Keck spectrum of component G at the same dispersion. The exposure time was also 900 s for component G. The vertical thinner grey dashed lines with labels at the lower right of the figure (6,608.2 Å, 6,666.7 Å, 7,231.0 Å, and 8,694.0 Å) represent the positions of absorption lines of the respective ions redshifted to $z = 0.680$ of Ca II H&K (3,933.7 Å and 3,968.5 Å), G-band (4,304.4 Å), and Mg I B-band (5,175.3 Å), respectively. There are no data below ~4,900 Å in the spectrum of component G.

Table 1 **Astrometry and photometry for SDSS J1004+4112**

Object	RA (h m s)*	Dec. (° ′ ″)*	g†	r†	i†	z†	$\Delta\theta\ddagger$
A	10 04 34.794	+41 12 39.29	18.67 ± 0.03	18.70 ± 0.02	18.46 ± 0.02	18.43 ± 0.05	3.73"
B	10 04 34.910	+41 12 42.79	19.05 ± 0.06	19.10 ± 0.06	18.86 ± 0.06	18.92 ± 0.06	0.00"
C	10 04 33.823	+41 12 34.82	19.71 ± 0.03	19.73 ± 0.02	19.36 ± 0.03	19.31 ± 0.07	14.62"
D	10 04 34.056	+41 12 48.95	20.67 ± 0.04	20.51 ± 0.04	20.05 ± 0.04	20.00 ± 0.13	11.44"
G	10 04 34.170	+41 12 43.66	22.11 ± 0.40	20.51 ± 0.13	19.54 ± 0.09	19.04 ± 0.21	8.44"

*RA, right ascension; dec., declination. RA and dec. are given in the J2000 system. These celestial coordinates were measured on the basis of the celestial coordinates of component B. The positional errors of components A, C and D (not including the absolute positional errors of component B) are 0.01" and that of component G is 0.05" per coordinate.

†g, r, i and z mean the magnitudes of each band.

‡Separation angles relative to component B.

Keck I telescope at the W. M. Keck Observatory. The results are shown in Fig. 1. All four components indeed show quasar-like features, with all emission lines giving a consistent redshift $z = 1.734 \pm 0.002$; the velocity differences of the quasar components are $\sim 100 \text{ km s}^{-1}$, comparable to the observational uncertainty. Although it is not obvious from Fig. 1, there are C IV absorption line systems at $z = 1.732$ in each of the four quasar spectra; this is an absorption system associated with the quasar itself, further supporting the lensing hypothesis: the four quasar images are from the same physical source. The differences in their spectra may be explained by the modest time-variability of the source quasar over $\sim 1 \text{ yr}$, the expected gravitational lensing time-delay²⁵ among those different images.

Additional strong support for the lensing hypothesis comes from the identification of the galaxy cluster responsible for the large-separation lensing. From the observed image separations (the maximum separation is 14.62", between images B and C), we infer that the lensing object should have a velocity dispersion in

excess of 600 km s^{-1} . Thus the lensing object cannot be a single galaxy, but must be rather a group or cluster of galaxies that has a sufficiently concentrated distribution of dark matter. To identify the lensing object, we obtained deep optical images of the system using the Subaru telescope of the National Astronomical Observatory, Japan. The result is shown in Fig. 2. A number of galaxies are clearly detected around component G, suggesting that it is the most luminous galaxy of the cluster. We obtained a spectrum of component G using the Keck I telescope. The spectrum shows a number of absorption features characteristic of an early-type galaxy at a

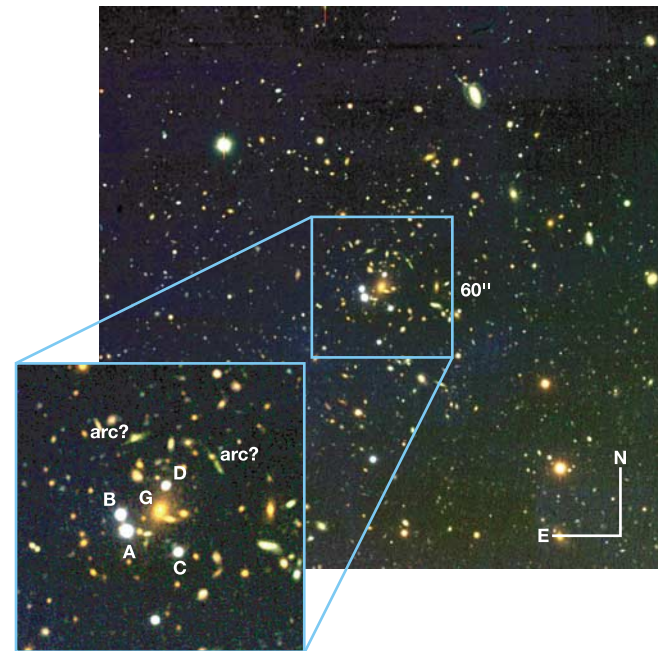


Figure 2 The *gri* composite Subaru image of the field around SDSS J1004+4112. The data were taken using the Subaru Prime Focus Camera³⁰ of the Subaru telescope. The magnitude limit is $i \approx 26.0$. The central 60" square is shown in an expanded view. The four quasar components are marked as A, B, C and D, and the bright galaxy located between the four quasar components is marked as G. The separation between components A and D is 12.77", and that between components B and C is 14.62". The positions (J2000) and the magnitudes of the components A–D and the brightest galaxy (component G) between the four quasar components are summarized in Table 1. Many faint galaxies can be seen—their positions and colours are consistent with being members of a cluster ($z = 0.68$) centred on component G. Two possible arclets (marked as 'arc?') can also be seen. The seeing had a full-width at half-maximum of 0.6".

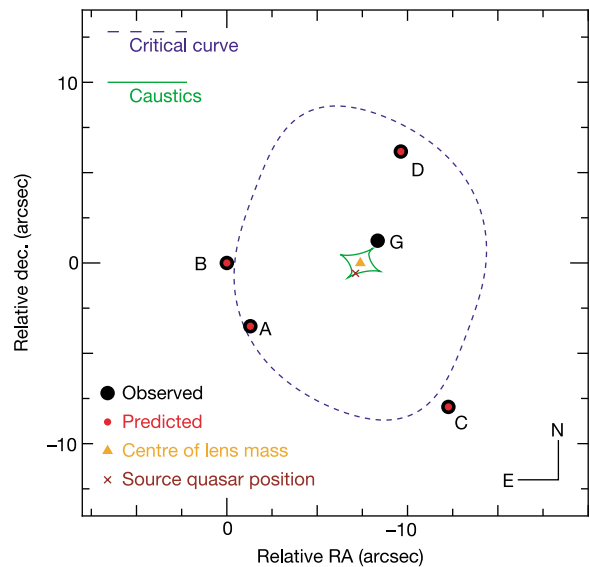


Figure 3 The best-fit lens model prediction compared with the observation. We used a lensing model of a singular isothermal ellipsoid (SIE) with external shear. The best-fit model has an Einstein radius of the SIE model $\alpha_E = 6.906''$ (corresponding to a velocity dispersion $\sim 700 \text{ km s}^{-1}$ at the cluster redshift), magnitude and position angle of the shear $\gamma = 0.250$ and $\theta_\gamma = -60.925^\circ$ (measured east of north), and ellipticity and its position angle $e = 0.498$ and $\theta_e = 21.434^\circ$ (measured east of north), with a source quasar position ($\Delta\text{RA}, \Delta\text{dec.}$) = $(-7.124'', -0.574'')$ and a centre of lensing mass ($\Delta\text{RA}, \Delta\text{dec.}$) = $(-7.387'', -0.004'')$ relative to the centre of component A. The black filled circles represent the observed positions of components A, B, C, D and G, and the red filled circles represent the predicted positions of components A–D. The green solid line is the position of the caustic in the source plane, and the blue dashed line represents the critical curve in the image plane. The small brown filled cross is the predicted position of the source quasar, and the small orange filled triangle is the predicted position of the centre of the lens mass. The differences between the observed and modelled image positions are much smaller than the observational uncertainties. The flux ratios predicted from the model, B/A, C/A and D/A are 0.78, 0.43 and 0.22, respectively. The total magnification of the quasar images predicted by the model is 56.48. The predicted flux ratios are close to the observational results; B/A = 0.69 ± 0.04 , C/A = 0.46 ± 0.02 and D/A = 0.25 ± 0.01 (measured from the i-band image). Microlensing by substructures and/or reddening by the Mg II absorption line systems that are seen in each spectrum might be the cause of the differences between the predicted flux ratios and the observations.

redshift of $z = 0.6799 \pm 0.0001$. We also obtained spectra of two faint galaxies immediately to the southwest of component G (Fig. 2) using the Faint Object Camera and Spectrograph²⁶ of the Subaru telescope. The redshifts of these two faint galaxies are $z = 0.6751 \pm 0.0001$, strongly suggesting a cluster of galaxies at $z \approx 0.68$ centred on component G. Clusters are dominated by elliptical galaxies, which all have very similar spectral energy distributions.

Many of the faint galaxies in Fig. 2 (~40 galaxies around component G) have colours that are similar to that of component G. The colours are consistent with the expected colours of elliptical galaxies at $z \approx 0.68$ ($g - r \approx 1.8$ and $r - i \approx 1.1$). In addition, there is an X-ray source in this direction detected by the ROSAT All-Sky Survey²⁷ (0.236 counts per second in a 473-s exposure). The emission, however, comes most probably from the quasar, because the detected X-ray flux is too strong for typical clusters of galaxies at $z = 0.68$. Finally, we note two possible arclets (highly distorted images of background galaxies due to gravitational lensing) in Fig. 2 (marked as 'arc?') close to component D. If future observations confirm that the arclets are indeed lensed background galaxies, they will provide strong additional constraints on the total mass distribution of the lensing cluster.

The lensing interpretation is further supported by a theoretical model of SDSS J1004+4112. We fitted the positions of the four quasar components with a singular isothermal ellipsoid (SIE) plus external shear model using lens modelling software²⁸. The best-fit model is illustrated in Fig. 3. The positions and relative brightnesses of all components agree well with the lens model predictions. The centre of the lensing mass is offset from the centre of component G by about 10 kpc at the cluster redshift, but the brightest galaxy of a cluster is not always found exactly at the centre of the potential well of that cluster.

The identical redshifts ($z = 1.734$) and the spectral energy distributions of the four lensed components, the existence of a lensing cluster of galaxies ($z = 0.68$), and the presence of possible arclets confirm the hypothesis that the quasar is lensed by this cluster. Furthermore, a theoretical lensing model involving the cluster and external shear simultaneously accounts for the observed geometry of the system and the relative brightness of the images. The present work represents the discovery of a long-predicted but previously undetected population of large-separation lensed quasars. □

Received 30 July; accepted 23 October 2003; doi:10.1038/nature02153.

- Narayan, R. & White, S. D. M. Gravitational lensing in a cold dark matter universe. *Mon. Not. R. Astron. Soc.* **231**, 97–103 (1988).
- Wambsganss, J., Cen, R., Ostriker, J. P. & Turner, E. L. Testing cosmogonic models with gravitational lensing. *Science* **268**, 274–276 (1995).
- Keeton, C. R. & Madau, P. Lensing constraints on the cores of massive dark matter halos. *Astrophys. J.* **549**, L25–L28 (2001).
- Wyithe, J. S. B., Turner, E. L. & Spergel, D. N. Gravitational lens statistics for generalized NFW profiles: Parameter degeneracy and implications for self-interacting cold dark matter. *Astrophys. J.* **555**, 504–523 (2001).
- Takahashi, R. & Chiba, T. Gravitational lens statistics and the density profile of dark halos. *Astrophys. J.* **563**, 489–496 (2001).
- Oguri, M. Constraints on the baryonic compression and implications for the fraction of dark halo lenses. *Astrophys. J.* **580**, 2–11 (2002).
- Navarro, J. F., Frenk, C. S. & White, S. D. M. A universal density profile from hierarchical clustering. *Astrophys. J.* **490**, 493–508 (1997).
- Maoz, D., Rix, H., Gal-Yam, A. & Gould, A. Survey for large-image separation lensed quasars. *Astrophys. J.* **486**, 75–84 (1997).
- Ofek, E. O., Maoz, D., Prada, F., Kolatt, T. & Rix, H. A survey for large-separation lensed FIRST quasars. *Mon. Not. R. Astron. Soc.* **324**, 463–472 (2001).
- Phillips, P. M. *et al.* The JVAS/CLASS search for 6-arcsec to 15-arcsec image separation lensing. *Mon. Not. R. Astron. Soc.* **328**, 1001–1015 (2001).
- Zhdanov, V. I. & Surdej, J. Quasar pairs with arcminute angular separations. *Astron. Astrophys.* **372**, 1–7 (2001).
- Kochanek, C. S. *et al.* CASTLES survey. (<http://cfa-www.harvard.edu/castles/>) (2003).
- Walsh, D., Carswell, R. F. & Weymann, R. J. 0957+561 A, B—Twin quasistellar objects or gravitational lens? *Nature* **279**, 381–384 (1979).
- Colley, W. N., Tyson, J. A. & Turner, E. L. Unlensing multiple arcs in 0024+1654: Reconstruction of the source image. *Astrophys. J.* **461**, L83–L86 (1996).
- Turner, E. L., Ostriker, J. P. & Gott, J. R. III The statistics of gravitational lenses—the distributions of image angular separations and lens redshifts. *Astrophys. J.* **284**, 1–22 (1984).

- York, D. G. *et al.* The Sloan Digital Sky Survey: Technical summary. *Astron. J.* **120**, 1579–1587 (2000).
- Stoughton, C. *et al.* Sloan Digital Sky Survey: Early data release. *Astron. J.* **123**, 485–548 (2002).
- Gunn, J. E. *et al.* The Sloan Digital Sky Survey photometric camera. *Astron. J.* **116**, 3040–3081 (1998).
- Pier, J. R. *et al.* Astrometric calibration of the Sloan Digital Sky Survey. *Astron. J.* **125**, 1559–1579 (2003).
- Hogg, D. W., Finkbeiner, D. P., Schlegel, D. J. & Gunn, J. E. A photometricity and extinction monitor at the Apache Point Observatory. *Astron. J.* **122**, 2129–2138 (2001).
- Smith, J. A. *et al.* The u'g'r'i'z' standard-star system. *Astron. J.* **123**, 2121–2144 (2002).
- Fukugita, M. *et al.* The Sloan Digital Sky Survey photometric system. *Astron. J.* **111**, 1748–1756 (1996).
- Blanton, M. R. *et al.* An efficient targeting strategy for multiobject spectrograph surveys: The Sloan Digital Sky Survey "tiling" algorithm. *Astron. J.* **125**, 2276–2286 (2003).
- Richards, G. T. *et al.* Spectroscopic target selection in the Sloan Digital Sky Survey: The quasar sample. *Astron. J.* **123**, 2945–2975 (2002).
- Oguri, M., Taruya, A., Suto, Y. & Turner, E. L. Strong gravitational lensing time delay statistics and the density profile of dark halos. *Astrophys. J.* **568**, 488–499 (2002).
- Kashikawa, N. *et al.* FOCAS: The Faint Object Camera and Spectrograph for the Subaru Telescope. *Publ. Astron. Soc. Jpn* **54**, 819–832 (2002).
- Cao, L., Wei, J.-Y. & Hu, J.-Y. High X-ray-to-optical flux ratio RASS-BSC sources. I. The optical identification. *Astron. Astrophys. Suppl.* **135**, 243–253 (1999).
- Keeton, C. R. Computational methods for gravitational lensing. Preprint at (<http://xxx.lanl.gov/astro-ph/0102340/>) (2001).
- Oke, J. B. *et al.* The Keck Low-Resolution Imaging Spectrometer. *Publ. Astron. Soc. Pacif.* **107**, 375–385 (1995).
- Miyazaki, S. *et al.* Subaru prime focus camera—Suprime-Cam. *Publ. Astron. Soc. Jpn* **54**, 833–853 (2002).

Acknowledgements Funding for the creation and distribution of the SDSS Archive has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Aeronautics and Space Administration, the National Science Foundation, the US Department of Energy, the Japanese Monbukagakusho, and the Max Planck Society. The SDSS website is <http://www.sdss.org/>. The SDSS is managed by the Astrophysical Research Consortium (ARC) for the Participating Institutions. The Participating Institutions are The University of Chicago, Fermilab, the Institute for Advanced Study, The Japan Participation Group, The Johns Hopkins University, Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, University of Pittsburgh, Princeton University, the United States Naval Observatory, and the University of Washington. This Letter is based in part on data collected at the Subaru telescope, which is operated by the National Astronomical Observatory of Japan, W. M. Keck Observatory, which is operated as a scientific partnership among the California Institute of Technology, the University of California, and the National Aeronautics and Space Administration, and the Apache Point Observatory (APO) 3.5-m telescope, which is owned and operated by the Astrophysical Research Consortium. Part of this work was performed under the auspices of the U.S. Department of Energy at the University of California Lawrence Livermore National Laboratory.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to N.I. (inada@utap.phys.s.u-tokyo.ac.jp).

Subatomic movements of a domain wall in the Peierls potential

K. S. Novoselov¹, A. K. Geim¹, S. V. Dubonos³, E. W. Hill² & I. V. Grigorieva¹

¹Department of Physics, ²Department of Computer Sciences, University of Manchester, Manchester M13 9PL, UK

³Institute for Microelectronics Technology, 142432 Chernogolovka, Russia

The discrete nature of crystal lattices plays a role in virtually every material property. But it is only when the size of entities hosted by a crystal becomes comparable to the lattice period—as occurs for dislocations^{1–3}, vortices in superconductors^{4–6} and domain walls^{7–9}—that this discreteness is manifest explicitly. The associated phenomena are usually described in terms of a background Peierls 'atomic washboard' energy potential, which was first introduced for the case of dislocation motion^{1,2} in the 1940s. This concept has subsequently been invoked in many situations to describe certain features in the bulk behaviour of materials, but has to date eluded direct detection and experimental scrutiny at a microscopic level. Here we report observations of

the motion of a single magnetic domain wall at the scale of the individual peaks and troughs of the atomic energy landscape. Our experiments reveal that domain walls can become trapped between crystalline planes, and that they propagate by distinct jumps that match the lattice periodicity. The jumps between valleys are found to involve unusual dynamics that shed light on the microscopic processes underlying domain-wall propagation. Such observations offer a means for probing experimentally the physics of topological defects in discrete lattices—a field rich in phenomena that have been subject to extensive theoretical study^{10–12}.

In magnetic materials, a domain wall (DW) has to pass through different spin configurations as it moves from one atomic plane to another^{7–14}. Figure 1a shows two principal spin configurations of a DW in a simple lattice—these configurations exhibit the maximum and minimum energy, and correspond to the centre of the wall lying respectively at and between atomic planes. This spatial variation of the wall's energy is generally referred to as the Peierls potential. Its amplitude depends critically on the ratio between the DW's width δ and the lattice periodicity d and, realistically, the Peierls potential is only observable for $\delta/d < 10$. For larger δ/d , the potential becomes so small ($\ll 1$ G) that pinning on defects should conceal it completely. The vast theoretical and experimental evidence gathered over several decades and based on studies of bulk properties of magnetic alloys has confirmed the existence of the magnetic Peierls potential, with probably the most definite conclusions drawn from

measurements of magnetic viscosity (for example, see refs 14–16). However, no experiment has yet been capable of directly detecting propagation through the magnetic (or any other) Peierls potential. In the present work, we revisit the Peierls potential by using a state-of-the-art technique of ballistic Hall micromagnetometry, previously used in studies of superconducting vortices^{17,18} and now adapted to the detection of DW movements. The approach allows us to resolve subatomic changes in the position of individual micrometre-sized segments of DWs and study their inter- and intra-Peierls valley movements.

For our studies, we have chosen thin films of yttrium-iron garnet, (YBi)₃(FeGa)₅O₁₂ (YIG), which combine relatively narrow walls ($\delta \approx 11$ nm at liquid-helium temperatures) with a large unit cell of size $a \approx 1.24$ nm, and provide $\delta/d \approx 6$ (Methods). Equally important is the high crystal quality of our samples^{19–21}, manifested in a coercivity of <0.1 G at room temperature and <10 G at liquid-helium temperatures, such that obscuring effects due to pinning on defects are relatively small. The YIG films have perpendicular magnetization and a domain structure shown in Fig. 1b for the case of room temperature. At low temperatures, the domain structure becomes pronouncedly triangular with long straight DWs. A submillimetre piece of the film was placed in immediate contact with the surface of a device consisting of micrometre-sized Hall sensors made from a two-dimensional electron gas (2DEG) following the microfabrication procedure described in refs 17 and 18 (Fig. 1b). Hall sensors in our experiments play the role of highly sensitive position detectors, which provide a spatial resolution of <1 Å with respect to DW movements. Their operation (explained in Methods) relies on the high sensitivity of such probes to changes in magnetic flux Φ induced by DW movements inside the central area of a Hall cross^{17,18,22–26}. For brevity, we discuss only experiments where the studied DWs were aligned parallel with the Hall device and moved in $\{110\}$ directions. In this well defined geometry, changes in the wall position Δx can be calculated directly from

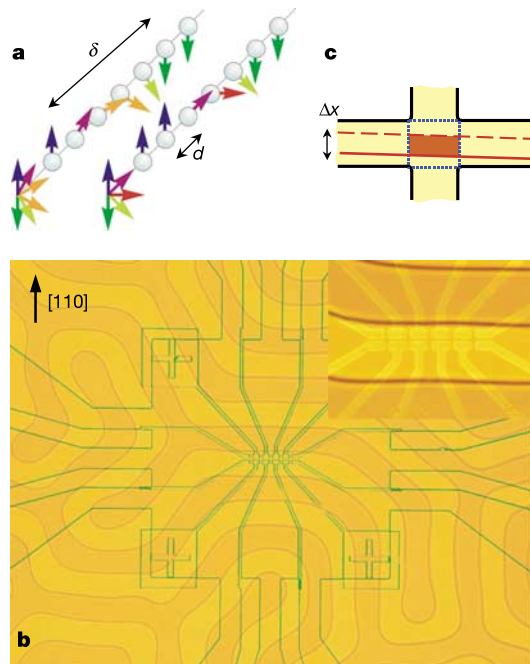


Figure 1 Experimental structures and devices. **a**, Principal spin configurations for a narrow Bloch wall: its centre either coincides with one of the atomic planes (right) or lies between them (left). The fan diagrams show the orientation of individual spins, looking in the direction perpendicular to the DW. **b**, The micrographs show a set of micrometre-sized Hall probes placed on top of a ferromagnetic garnet. Edges of the 2DEG mesa are seen as green lines on the micrograph. The image overlays a photograph of the domain structure taken in transmitted polarized light at room temperature. The inset magnifies the central part of the experimental structure. The scale is given by the domains' width of $\sim 14 \mu\text{m}$ and the size of Hall probes ($1.5 \mu\text{m}$). By measuring simultaneously the response at different Hall crosses, we ensured that at low temperatures the studied DWs were parallel to the set of sensors as the photograph shows and moved as rigid planes (Methods). **c**, The drawing illustrates that a shift in the average position of a wall Δx induces a change in flux $\Delta \Phi$ inside the sensitive area marked by the dotted lines^{17,22}. This leads to a linear change in Hall resistance, which was recorded in the experiments.

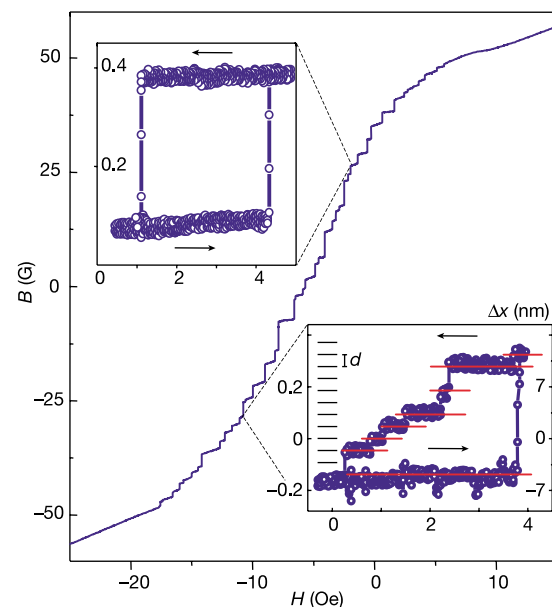


Figure 2 Nanometre movements of domain walls over submicrometre distances. Main panel, a typical Hall response measured by a $1.5\text{-}\mu\text{m}$ Hall cross as a domain wall slowly creeps from one of its sides to the other ($T = 0.5$ K). For convenience, the Hall response is given in terms of the average local field B inside the cross, which is calculated by using the measured Hall coefficient. The insets show examples of local hysteresis loops; abscissa, ΔH (Oe); left-hand ordinate, ΔB (G); right-hand ordinate (lower inset only), Δx (nm). The red lines are guides to the eyes, indicating constant positions of the DW.

the measured Hall signal, without using any unknown parameters (Methods).

To move a DW, we slowly varied the external field H applied perpendicular to the garnet film. Figure 2 shows a typical example of changes in local field B detected by a Hall sensor as a DW crosses it from one side to the other. We see that B changes its sign, which reflects the change in polarity of the domain above the sensor, and $B = 0$ corresponds to the state where the wall lies exactly in the middle. The overall shape of the transition curve is in good agreement with a simple theory²¹. Overlaid on this universal behaviour, we see a number of small sample- and sweep-dependent steps, indicating that a DW does not move smoothly but covers micrometre-long distances in a series of small jumps. Such jumps have previously been studied by several techniques (for example, refs 27–29), and are usually referred to as Barkhausen noise. A typical step in Fig. 2 corresponds to a wall moving by 10–50 nm. While a DW was located within the Hall cross, we could reverse a field sweep to investigate the local coercivity of the wall (left inset in Fig. 2). Such hysteresis loops are usually reproducible for many field cycles, and we attribute them to pinning on individual defects^{21,28}.

In addition to the above behaviour, the high resolution of the 2DEG micromagnetometry allowed us to discern very small DW jumps (right inset in Fig. 2), which stood out from the ‘ordinary’ jumps for two reasons. First, they matched closely the lattice periodicity in the direction of DW travel $\langle 110 \rangle$ ($d = 2^{1/2}a \approx 1.75$ nm) and, second, they were practically the only jumps observed in the range below ~ 10 nm. The use of statistical analysis techniques (which is standard practice in, for example, particle physics) shows that—with a confidence level of 94%—the data shown in the right inset of Fig. 2 correspond to an event comprising several steps of equal length (four single and three double steps), where the length of a single step is d . To obtain further proof that such steps indeed reveal jumps between equivalent crystal lattice positions, we carried out the complementary experiments described below.

As a DW moves through a crystal, it interacts with a large number of pinning sites and becomes bent and strained in the process. We can generally expect that a strained wall would jump between strong pinning sites without being affected by the weaker ones. This is

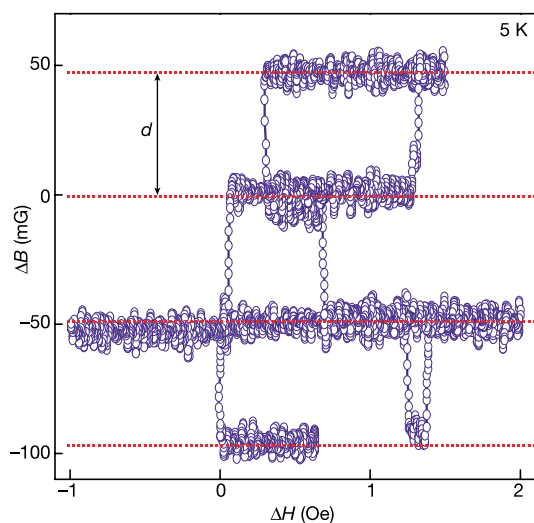


Figure 3 Jumps of a domain wall between equivalent lattice sites. The data were taken during a very slow sweep (~ 1 h), which was required to achieve the subatomic resolution. For such time intervals, relaxation processes lead to irreproducible changes in the domain structure, usually far away from the detection site (this can be seen as occasional DW jumps at a constant H). On the graph, this results in the same position of a DW for different values of H . For clarity, we subtracted a small smooth background in B associated with changes in the local stray field induced by other domains.

clearly seen by magnetic force microscopy (at room temperature). To release the strain, we demagnetized the sample by applying an a.c. magnetic field with an amplitude h gradually decreasing from ~ 5 G to zero, while a constant field H kept the wall close to the centre of the probe. This proved to be a critical improvement: in the demagnetized state, DWs started to propagate via clear quantized jumps matching the lattice periodicity. The distance between the equivalent sites was measured to be 1.6 ± 0.2 nm, in agreement with the Peierls potential periodicity $d = 1.75$ nm. Figure 3 shows an example of such behaviour, which leaves no doubt of the presence of a periodic atomic landscape impeding DW movements.

From Fig. 3, we can also estimate that it requires a field of ~ 1 G to move a DW out of a well created by adjacent crystal planes (that is, intrinsic pinning is several times weaker than pinning on a typical defect in our garnets; see Fig. 2). Further experiments yielded a value of the intrinsic coercive field $H_C \approx 0.7$ G at $T < 10$ K. The theory of the magnetic Peierls potential predicts H_C to be of the order of^{13–15}

$$H_C \approx C(A/d^2 M_S) \exp(-\pi \delta/d)$$

where $C \approx 10^3$ (ref. 13) and M_S is the saturation magnetization. Taking into account the exponential dependence of H_C on A and K (Methods), which are known to $\sim 10\%$ accuracy, the formula yields H_C in the range from 0.1 to 5 G, in agreement with the experiment.

In addition to the detection of the Peierls potential, we have studied its shape, which is predicted to be sinusoidal^{7,13}. This prediction implies that a DW should remain somewhat mobile within Peierls valleys—that is, not pinned completely. Such intra-valley movements are expected to be ~ 1 Å and, therefore, could not be resolved in our d.c. magnetization data (compare Fig. 3). To gain information about the finest DW movements, we measured local a.c. susceptibility χ (a.c. measurements provide a higher flux sensitivity, and hence a higher spatial resolution). To this end, in addition to the d.c. field H that controls DW's position, we applied an oscillating field h and measured an a.c. signal generated by oscillatory movements of a DW. Changes in χ show how the mobility of a DW varies with its position inside a Peierls valley. Using this approach, we confirmed that a domain wall could indeed move near the bottom of a valley, and the detected a.c. signal corresponded to an average shift of a DW by up to ~ 0.5 Å. In addition to this, however, a.c. measurements revealed strikingly

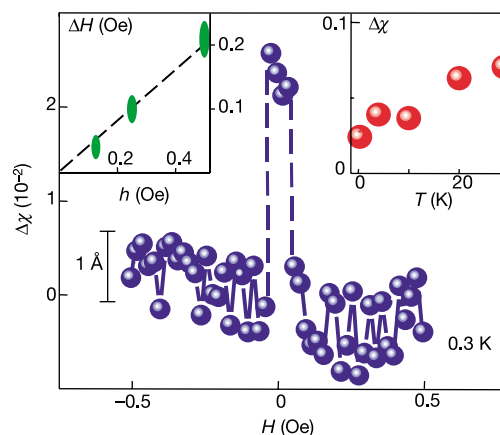


Figure 4 Domain wall on the Peierls ridge. Main panel, changes in local a.c. susceptibility χ while d.c. field H moves the wall from one Peierls valley to the next (a.c. field has amplitude 0.5 G and frequency 8 Hz). We subtracted a constant background due to the Hall response induced directly by the a.c. field. The slight variation of χ seen on the curve away from the transient state is not reproducible for different walls and after thermal cycling. The insets show the dependence of the width ΔH of the transient state on h , and the dependence of its amplitude $\Delta\chi$ on temperature.

unusual DW dynamics, in qualitative disagreement with the behaviour expected for an object moving in a tilted sinusoidal potential.

One of the most notable features that we observed is a large, well-reproduced peak in DW mobility (Fig. 4). Here, zero H corresponds to a DW position in the middle between two adjacent Peierls valleys, as simultaneously detected in the d.c. magnetization signal (the latter shows a smeared transition between two DW positions separated by d). The peak has abrupt edges (shown by dashed lines in Fig. 4); that is, above a certain value of H the oscillating wall suddenly falls into a neighbouring Peierls valley and becomes locked there. When h was switched off for a few seconds at a constant H close to one of the peak's edges, the transient mobile state did not recover on switching the modulation back. This indicates that it requires a time of ~ 1 s for the wall to become locked in a Peierls valley. With decreasing h from 0.5 to 0.1 G, both the width and the amplitude of the peak shrink linearly (inset in Fig. 4). The observed behaviour suggests that an a.c. field stabilizes a DW in what should be an intrinsically unstable state between two Peierls valleys.

We can interpret the transient state as the centre of a DW sitting effectively on a Peierls ridge, kept there by an oscillating magnetic force. This situation closely resembles the so-called reversed pendulum, which can be stabilized in the unstable upside-down position by an oscillating force^{30,31}. This analogy allows us to describe the observed resonance semiquantitatively but does not provide a microscopic picture. To this end, we invoke the well-known 'kink' model^{2,14}, where a DW moves between Peierls valleys via a process where at first only a submicrometre segment of a DW (a jog) moves to the next valley. Spreading the boundary of such a jog along the wall eventually leads to the relocation of the whole DW. It is plausible that a.c. modulation could stabilize the jog so that its boundaries move back and forth inside the sensitive area of a Hall cross without collapsing, until changes in H extend the jog outside this area, where it eventually becomes pinned. Indeed, the maximum value of $\Delta\chi$ (observed at 30 K and 0.5 G) corresponds to movements of a DW by $\Delta x \approx 1$ nm; that is, as if nearly the whole segment of the wall inside the Hall cross swings between adjacent valleys.

The single-jog model provides a sensible description for the behaviour shown in Fig. 4, as well as for the majority of other a.c. susceptibility results (to be reported elsewhere). However, the origin of the long characteristic times remains puzzling. Moreover, we believe that the kink/jog picture may no longer be justifiable for the case of $\Delta x \ll \delta$ as, on this scale, the spin configuration of a DW changes (the DW can 'breathe') and we cannot simply refer to an average shift of a DW as a whole. We believe that the detected transient state indicates some internal modes being excited inside a DW when it is 'softened' and ready to move from one valley to another.

Further experimental and theoretical work is required to understand atomic-scale DW dynamics. This physics has been extensively discussed in theory (for example, solitons on discrete lattices^{10–12}), but has not until now been accessible in a direct experiment to allow the testing of a variety of theoretical models. The present results demonstrate the possibility of studying the physics of domain walls at a new level of experimental resolution, and in a controlled and reproducible manner. This development is likely to lead not only to refinements of the existing models, but also to a greater depth of understanding of fundamental and technologically important phenomena governed by movements of domain walls. □

Methods

Samples

We studied a 10- μm -thick $(\text{YBi})_3(\text{FeGa})_5\text{O}_{12}$ film grown in the $[111]$ direction. The material has a saturation magnetization $4\pi M_s \approx 200$ G, an exchange energy $A \approx 1.8 \times 10^{-7}$ erg cm^{-1} and, below 10 K, its crystal anisotropy $K \approx 1.4 \times 10^6$ erg cm^{-3} yields a wall thickness $\delta = \pi(A/K)^{1/2} \approx 11$ nm. The above values were found with accuracy $\sim 10\%$.

Owing to in-plane crystal anisotropy, DWs tend to lie in three equivalent planes, $\{110\}$.

This crystallographic alignment is already seen at 300 K (Fig. 1b), and becomes stronger at lower temperatures as the anisotropy increases. There, domain walls become straight over distances of ~ 1 mm. Using alignment marks, we placed our 2DEG sensors inside a chosen area of a YIG film with many parallel domains and aligned the sensors parallel to them, that is, perpendicular to one of $\{110\}$ axes. The spacing between the garnet and 2DEG was measured to be less than 100 nm (ref. 21). We restricted the reported experiments to temperatures below 30 K, mainly because of thermally activated relaxation processes, which led to irreproducible changes in the domain structure and did not allow accurate measurements that require slow sweeps of magnetic field.

There are two periodic sets of equivalent crystallographic positions for a $\{110\}$ DW, which are separated by b and $2b$ (where $b = a/\sqrt{2}$ is the distance between the nearest basal planes). They require the translation of wall's spin configuration in directions (001) and (110) , respectively, and involve different exchange interactions¹³. Both periods should contribute to the Peierls potential, but because of the exponential dependence on δ/d only the longest periodicity $d = 2a/\sqrt{2} \approx 1.75$ nm can be expected to remain observable. Our measurements did find this periodicity, but it remains unclear why a DW could not avoid the observed Peierls barriers by exploiting 'the third dimension' (that is, moving by two smaller oblique jumps in (001) rather than by the straight jumps perpendicular to the DW's plane). We finally note that our analysis ignores the complex unit cell structure of garnets, which may also play some role.

Micromagnetometers

Submicrometre Hall probes made from a 2DEG were used as position detectors of a domain wall. When a DW enters the sensitive area of a probe, its response R_H starts changing. A shift Δx in the DW's position leads to a change in magnetic field B and flux Φ through the Hall cross (Fig. 1c), which in turn induces a Hall response such that $\Delta R_H = \alpha \Delta \Phi = \beta \Delta x$ (refs 17, 22). The second part of the equation assumes that a DW is straight within the sensitive area of the Hall cross. This is justified for the reported experiments because we could simultaneously measure R_H at different crosses and, at low T , found a nearly perfect correlation between movements of DWs detected by neighbouring Hall crosses²¹. This indicates that DWs move as fairly rigid objects, so that their large segments (up to $10 \mu\text{m}$ in size for $T < 10$ K) shift as a whole (that is, without bending). α and β are found experimentally^{17,21,22} and, therefore, changes in R_H can be translated in DW movements without any fitting parameters.

We used Hall sensors made from a high-mobility 2DEG because of their exceptional sensitivity to flux variations $\Delta \Phi$ on a submicrometre scale¹⁷. At temperatures $T < 80$ K, such sensors effectively work as fluxmeters and are capable of resolving $\Delta \Phi \approx 10^{-4} \phi_0$, where ϕ_0 is a flux quantum. This technique has previously been used and fully described in studies of submicrometre superconducting^{17,18} and ferromagnetic^{23–26} particles. In the context of the present work, we have exploited this unique flux sensitivity of 2DEG sensors to achieve a subatomic resolution in the average position of individual DWs. Indeed, if a DW passes the whole width w of a cross, Φ changes by several ϕ_0 (for the given value of M_s in our garnets). On the other hand, as we can resolve $\Delta \Phi \approx 10^{-4} \phi_0$, this corresponds to a shift of a DW by $\Delta x \approx w(\Delta \Phi/\phi_0) \approx 1$ Å. Note that many magnetic materials have larger values of M_s and, hence, the magnetometers should provide even higher spatial resolution (< 0.1 Å). We mainly used 2DEG crosses with $w \approx 1$ – $2 \mu\text{m}$ because, in our experience, they exhibit low noise and allow the maximum spatial resolution.

Received 7 July; accepted 29 October 2003; doi:10.1038/nature02180.

1. Peierls, R. E. The size of a dislocation. *Proc. Phys. Soc.* **52**, 34–37 (1940).
2. Nabarro, F. R. N. Dislocations in a simple cubic lattice. *Proc. Phys. Soc.* **59**, 256–272 (1947).
3. Kolar, H. R., Spence, J. C. H. & Alexander, H. Observation of moving dislocation kinks and unpinning. *Phys. Rev. Lett.* **77**, 4031–4034 (1996).
4. Ivlev, B. I. & Kopnin, N. B. Flux creep and flux pinning in layered high-temperature superconductors. *Phys. Rev. Lett.* **64**, 1828–1830 (1990).
5. Oussena, M., deGroot, P. A. J., Gagnon, R. & Taillefer, L. Lock-in oscillations in magnetic hysteresis curves of $\text{YBa}_2\text{Cu}_3\text{O}_{7-x}$ single crystals. *Phys. Rev. Lett.* **72**, 3606–3609 (1994).
6. Kleiner, R., Steinmeyer, F., Kunkel, G. & Müller, P. Intrinsic Josephson effects in $\text{Bi}_2\text{Sr}_2\text{CaCu}_2\text{O}_8$ single crystals. *Phys. Rev. Lett.* **68**, 2394–2397 (1992).
7. Barbara, B. Propriétés des parois étroites dans les substances ferromagnétiques à forte anisotropie. *J. Phys.* **34**, 1039–1046 (1973).
8. van den Broek, J. J. & Zijlstra, H. Calculation of intrinsic coercivity of magnetic domain walls in perfect crystals. *IEEE Trans. Magn.* **7**, 226–230 (1971).
9. Egami, T. & Graham, C. D. Jr Domain walls in ferromagnetic Dy and Tb. *J. Appl. Phys.* **42**, 1299–1300 (1971).
10. Aubry, S. Devils staircase and order without periodicity in classical condensed matter. *J. Phys.* **44**, 147–162 (1983).
11. Braun, O. M. & Kivshar, Y. S. Nonlinear dynamics of the Frenkel-Kontorova model. *Phys. Rep.* **306**, 2–108 (1998).
12. Bishop, A. R. & Lewis, W. F. A theory of intrinsic coercivity in narrow domain wall materials. *J. Phys. C* **12**, 3811–3825 (1979).
13. Hilzinger, H. R. & Kronmüller, H. Spin configuration and intrinsic coercive field of narrow domain walls in Co_2R -compounds. *Phys. Status Solidi* **54**, 593–604 (1972).
14. Egami, T. Theory of intrinsic magnetic after-effect. *Phys. Status Solidi* **19**, 747–758 (1973).
15. Arnaud, J. I., del Moral, A. & Abell, J. S. Intrinsic coercive field in pseudobinary cubic intermetallic compounds. *J. Magn. Magn. Mater.* **61**, 370–380 (1986).
16. Musznikov, N. V., Andreev, A. V., Korolyov, A. V. & Shiohara, Y. Magnetic viscosity in a UPTAl single crystal. *J. Alloys Comp.* **305**, 188–193 (2002).
17. Geim, A. K. et al. Ballistic Hall micromagnetometry. *Appl. Phys. Lett.* **71**, 2379–2381 (1997).
18. Geim, A. K. et al. Phase transitions in individual sub-micrometre superconductors. *Nature* **390**, 259–262 (1997).
19. Pisarev, R. V. et al. Optical second-harmonic generation in magnetic garnet thin films. *J. Phys.* **5**, 8621–8628 (1993).

20. Pavlov, V. V., Pisarev, R. V., Kirilyuk, A. & Rasing, T. Observation of a transversal nonlinear magneto-optical effect in thin magnetic garnet films. *Phys. Rev. Lett.* **78**, 2004–2007 (1997).
21. Novoselov, K. S., Geim, A. K., van der Berg, D., Dubonos, S. V. & Maan, J. C. Domain wall propagation on nanometre scale: coercivity of a single pinning center. *IEEE Trans. Magn.* **38**, 2583–2585 (2002).
22. Peeters, F. M. & Li, X. Q. Hall magnetometer in the ballistic regime. *Appl. Phys. Lett.* **72**, 572–574 (1998).
23. Kent, A. D., von Molnar, S., Gider, S. & Awschalom, D. D. Properties and measurement of scanning tunneling microscope fabricated ferromagnetic particle arrays. *J. Appl. Phys.* **76**, 6656–6660 (1994).
24. Li, Y. Q. *et al.* Hall magnetometry on a single iron nanoparticle. *Appl. Phys. Lett.* **80**, 4644–4646 (2002).
25. Hengstmann, T. M., Grundler, D., Heyn, C. & Heitmann, D. Stray-field investigation on permalloy nanodisks. *J. Appl. Phys.* **90**, 6542–6544 (2001).
26. Schuh, D., Biberger, J., Bauer, A., Breuer, W. & Weiss, D. Hall-magnetometry on ferromagnetic dots and dot arrays. *IEEE Trans. Magn.* **37**, 2091–2093 (2001).
27. Vergne, R., Cotillard, J. C. & Porteseil, J. L. Some statistical aspects of magnetization processes in ferromagnetic bodies—motion of a single 180-degrees Bloch wall in an imperfect crystalline medium. *Rev. Phys. Appl.* **16**, 449–476 (1981).
28. Wunderlich, J. *et al.* Influence of geometry on domain wall propagation in a mesoscopic wire. *IEEE Trans. Magn.* **37**, 2104–2107 (2001).
29. Kim, D. H., Choe, S. B. & Shin, S. C. Direct observation of Barkhausen avalanche in Co thin films. *Phys. Rev. Lett.* **90**, 087203 (2003).
30. Magnus, K. *Vibrations* (Blackie & Son, London, 1965).
31. Acheson, D. *From Calculus to Chaos* (Oxford Univ. Press, Oxford, 1997).

Acknowledgements This research was supported by the EPSRC (UK). We thank S. Gillott and M. Sellers for technical assistance and J. Steeds for advice on dislocation motion. S.V.D. also acknowledges support from Russian Ministry of Science and Technology.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.K.G. (geim@man.ac.uk).

Subwavelength-diameter silica wires for low-loss optical wave guiding

Limin Tong^{1,2}, Rafael R. Gattass¹, Jonathan B. Ashcom^{1*}, Sailing He², Jingyi Lou², Mengyan Shen^{1,3}, Iva Maxwell¹ & Eric Mazur¹

¹Department of Physics and Division of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA

²Centre for Optical and Electromagnetic Research and Department of Physics, Zhejiang University, Hangzhou 310027, China

³Department of Physics, Graduate School of Science, Tohoku University, Sendai, Miyagi 9808578, Japan

* Present address: Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, Massachusetts 02420, USA

Silica waveguides with diameters larger than the wavelength of transmitted light are widely used in optical communications, sensors and other applications^{1–3}. Minimizing the width of the waveguides is desirable for photonic device applications, but the fabrication of low-loss optical waveguides with subwavelength diameters remains challenging because of strict requirements on surface roughness and diameter uniformity^{4–7}. Here we report the fabrication of subwavelength-diameter silica ‘wires’ for use as low-loss optical waveguides within the visible to near-infrared spectral range. We use a two-step drawing process to fabricate long free-standing silica wires with diameters down to 50 nm that show surface smoothness at the atomic level together with uniformity of diameter. Light can be launched into these wires by optical evanescent coupling. The wires allow single-mode operation, and have an optical loss of less than 0.1 dB mm^{−1}. We believe that these wires provide promising building blocks for future microphotonic devices with subwavelength-width structures.

The fabrication of thin silica wires was first investigated in the

nineteenth century, when the mechanical properties of the wires were studied, but their optical properties and applications remained uninvestigated^{8,9}. It was not until a century later that researchers began to investigate the optical applications of silica wires made by drawing high-purity glass fibres from a laser-heated melt^{10–14}. With a diameter of more than one micrometre, these silica wires allow multimode waveguiding of visible and infrared light. Submicrometre wires allow single-mode operation, but both theoretical and experimental results show that the laser power required for drawing silica submicrometre- or nanometre-diameter wires (SMNWs) with a uniform diameter is impractically large^{14,15}. When drawing wires from a flame-heated melt, turbulence and convection make it difficult to control the temperature gradient in the drawing region, and consequently size uniformity is difficult to maintain when the wire diameter is reduced to less than one micrometre. Silica nanowires with diameters ranging from ten to several hundred nanometres have recently been obtained with other methods^{16–18}, but the diameter fluctuation and sidewall roughness of those wires makes them unsuitable for low-loss optical wave guiding.

Here we introduce a two-step drawing process for fabricating long uniform silica SMNWs by a flame-heated fibre drawing method. First, we use a flame to draw a silica fibre to a micrometre-diameter wire. Second, to obtain a steady temperature distribution in the drawing region while further reducing the wire diameter, we use a tapered sapphire fibre with a tip diameter of about 80 μm to absorb the thermal energy from the flame (Fig. 1). The sapphire fibre taper, which is fabricated using laser-heating growth method¹⁹, confines the heating to a small volume and helps maintain a steady temperature distribution during the drawing. One end of a micrometre-diameter silica wire is placed horizontally on the sapphire tip, and the flame is adjusted until the temperature of the tip is just above the drawing temperature (about 2,000 K). We then rotate the sapphire tip around its axis of symmetry to wind the silica wire around the tip. The wire coil is moved about 0.5 mm out of the flame to prevent melting and the wire is then drawn perpendicular to the axis of the sapphire tip in the horizontal plane at a speed of 1–10 mm s^{−1} to form a SMNW.

Using this technique, we obtained silica SMNWs with diameters down to 50 nm and lengths up to tens of millimetres. Figure 2a shows a scanning electron microscope (SEM) image of a 4-mm-long wire with a diameter of 260 nm; the wire is roughly coiled up to show its length. The maximum diameter variation ΔD is about 8 nm over the 4-mm length L of the wire, giving $\Delta D/L = 2 \times 10^{-6}$. The excellent uniformity of wires with diameters ranging from 50 to 1,100 nm can also be seen in Fig. 2b–d. Higher-magnification transmission electron microscope (TEM) images of a 240-nm-

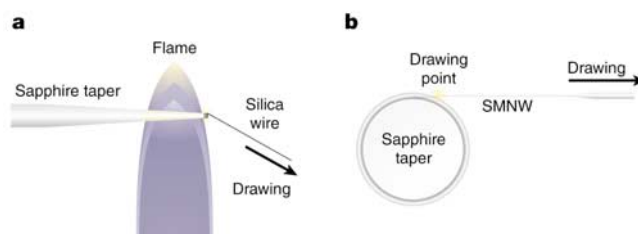


Figure 1 The second step in the fabrication process of silica submicrometre- and nanometre wires (SMNWs). **a**, Schematic diagram of the drawing of the wire from a coil of micrometre-diameter silica wire wound around the tip of a sapphire taper. The sapphire taper is heated with a CH₃OH torch with a nozzle diameter of about 6 mm. The wire is drawn in a direction perpendicular to the sapphire taper. **b**, Magnified view of the drawing process. The sapphire taper ensures that the temperature distribution in the drawing region remains steady.

diameter wire and of the sidewall of a 330-nm-diameter wire (Fig. 2e and f) show no visible irregularity in the surface of the wires. The typical sidewall root-mean-square roughness of these wires is less than 0.5 nm. The electron microscope pictures in Fig. 2 show that the wires reported here have much better uniformity of diameter and surface smoothness than submicrometre- or nanometre-width wires, strips or other structures obtained by other methods^{16–18,20–23}.

Because of their length and flexibility, the silica SMNWs can readily be manipulated under an optical microscope. For example, Fig. 3a shows a 520-nm-diameter silica wire that is elastically bent to form a microscopic ring with a diameter of less than 15 μm ; Fig. 3b shows two 330-nm-diameter wires that are twisted together. The wires do not break when bent and/or twisted, indicating that they have excellent flexibility and mechanical properties. Using the Young's modulus of silica (73.1 GPa), we find that the tensile strength of the wire in Fig. 3a is at least 2.5 GPa (refs 24, 25). The wires can be bent much more sharply than shown in Fig. 3a. Figure 3c, for example, shows a 280-nm-diameter wire bent to a radius of 2.7 μm , indicating a tensile strength exceeding 4.5 GPa. By bending them to the point of fracture, we find that the tensile strength of the wires is typically higher than 5.5 GPa. Using plastic bending, we have achieved bending radii of less than 1 μm .

We investigated the optical properties of the silica SMNW by sending light into them using evanescent coupling. First, a SMNW is suspended in air, with one end fixed to a support and the other end connected to the fibre taper from which it was drawn (fibre taper 1 in Fig. 4a). Light is then coupled into the SMNW through a second fibre taper. The launching wire from this second taper attaches itself to the guiding wire because of a van der Waals attraction between the two. To reduce the contribution from

scattered light, both tapers are gold-coated except for the region used for evanescent coupling. Figure 4b shows an optical micrograph of the coupling between a 390-nm-diameter launching wire and a 450-nm-diameter SMNW; Fig. 4c shows a 360-nm-diameter wire guiding light of 633-nm wavelength from the left. The light is intercepted at the right by a supporting 3- μm wire to show that the amount of light scattered by the wire is small compared to that guided by it. In Fig. 4d, a 550-nm-diameter wire guides 633-nm light in air (on the left) and along the surface of a MgF_2 crystal (on the right). Because the refractive index of MgF_2 is lower than that of silica (1.39 versus 1.46), the light continues to be guided by the wire on the MgF_2 surface, demonstrating the possibility of integrating SMNWs with low-index substrates for device applications. We also determined that a 620-nm-diameter wire immersed in water guides light, demonstrating the possibility of using these wires for chemical and biosensing in liquid media.

We determined the optical loss of the SMNWs by measuring their transmission as a function of the length L between the coupling region and taper 1 (see Fig. 4a). To maintain the same coupling efficiency between the launching and guiding wires, we adjust the overlap of the two wires until the output from fibre taper 1 is maximized. Figure 4e shows the optical loss at wavelengths of 633 and 1,550 nm for various SMNW diameters. At 633 nm, the optical

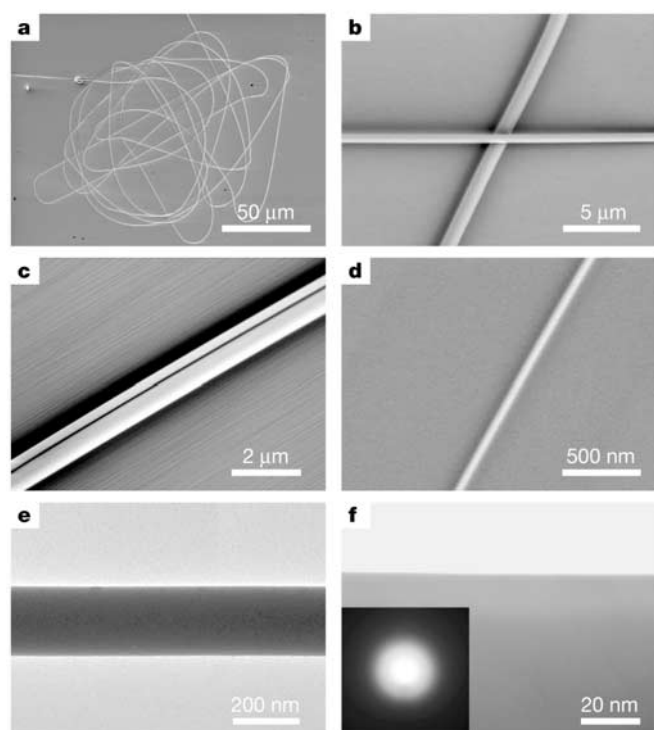


Figure 2 Electron micrographs of SMNWs. **a–d**, SEM images; **e,f**, TEM images. **a**, A coiled 260-nm-diameter silica wire with a total length of about 4 mm. **b**, Two crossed 570-nm and 1,100-nm diameter wires. **c**, Two parallel 170-nm and 400-nm diameter wires. **d**, A silica nanowire with a diameter of about 50 nm. **e**, A 240-nm-diameter silica wire. **f**, The surface of a 330-nm-diameter silica wire; the electron diffraction pattern (inset) demonstrates that the wire is amorphous.

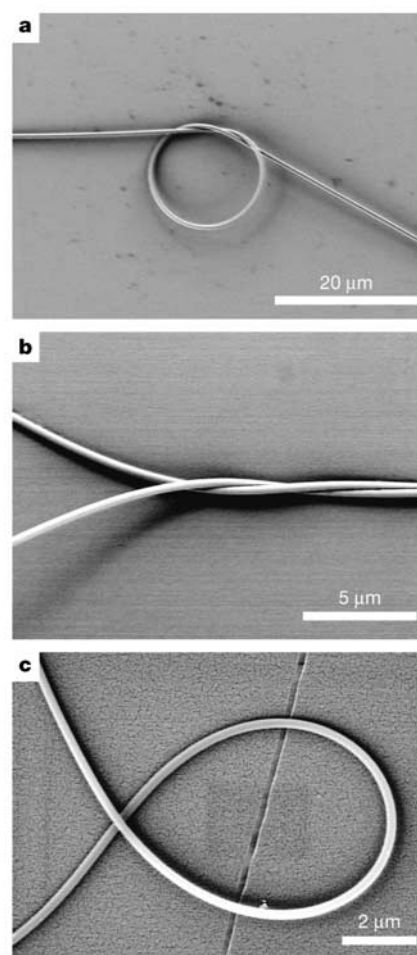


Figure 3 Micromanipulation and flexibility of SMNWs. **a**, SEM image of a 15- μm -diameter micro-ring made with a 520-nm-diameter silica wire. **b**, SEM image of two twisted 330-nm-diameter wires. **c**, SEM image of a bent 280-nm-diameter silica wire with a bending radius of 2.7 μm .

loss of a 190-nm-diameter wire is about 1.7 dB mm^{-1} , which is much lower than the optical loss of other subwavelength-structures such as metallic plasmon waveguides^{26–28}.

The increasing loss with decreasing wire diameter can be attributed to surface contamination: as the wire diameter is reduced below the wavelength, more light is guided outside the wire as an evanescent wave and becomes susceptible to scattering by surface contamination. Calculations show that the critical diameter for single-mode operation of a silica wire is about 450 nm at 633-nm wavelength and 1,100 nm at 1,550-nm wavelength²⁹, and that about 20% of the energy propagates outside the silica core under those conditions. Figure 4e shows that wires with these diameters have an optical loss below 0.1 dB mm^{-1} . For smaller diameters, more light propagates outside the silica core as an evanescent wave. Evanescent wave propagation is extremely useful for enhancing the performance of devices such as optical sensors.

Because of the large index contrast between silica and air, silica SMNWs can be bent sharply without incurring large bending losses. Using a three-dimensional finite-difference time domain

simulation³⁰, we find a bending loss of less than 0.3 dB for a 90° turn with a bending radius of $5 \mu\text{m}$ in an air-clad 450-nm-diameter silica wire (for light of 633-nm wavelength). These results indicate that the silica SMNWs are also suitable for applications where tight waveguide bends are desired. Using two probes from a scanning tunnelling microscope for micromanipulation, we were able to successfully guide 633-nm wavelength light through a bend with a $5.6\text{-}\mu\text{m}$ radius in a silica wire with a diameter of about 510 nm (see Fig. 4f). As can be seen in the figure, the intensity of scattered light after the bend is not greatly reduced, indicating that the bending loss is low. The ability to guide light through sharp bends is especially useful for miniaturization of photonic devices. Also, with low bending loss, micro-rings made from these wires (such as the one shown in Fig. 3a) can be incorporated into photonic devices such as optical microresonators for optical communication or optical sensing. For example, we used a 950-nm-diameter SMNW to make a ring with a $75\text{-}\mu\text{m}$ radius similar to the one shown in Fig. 3a; preliminary measurements show that the ring has Q factor of about 1,500 at a wavelength of 1,550 nm.

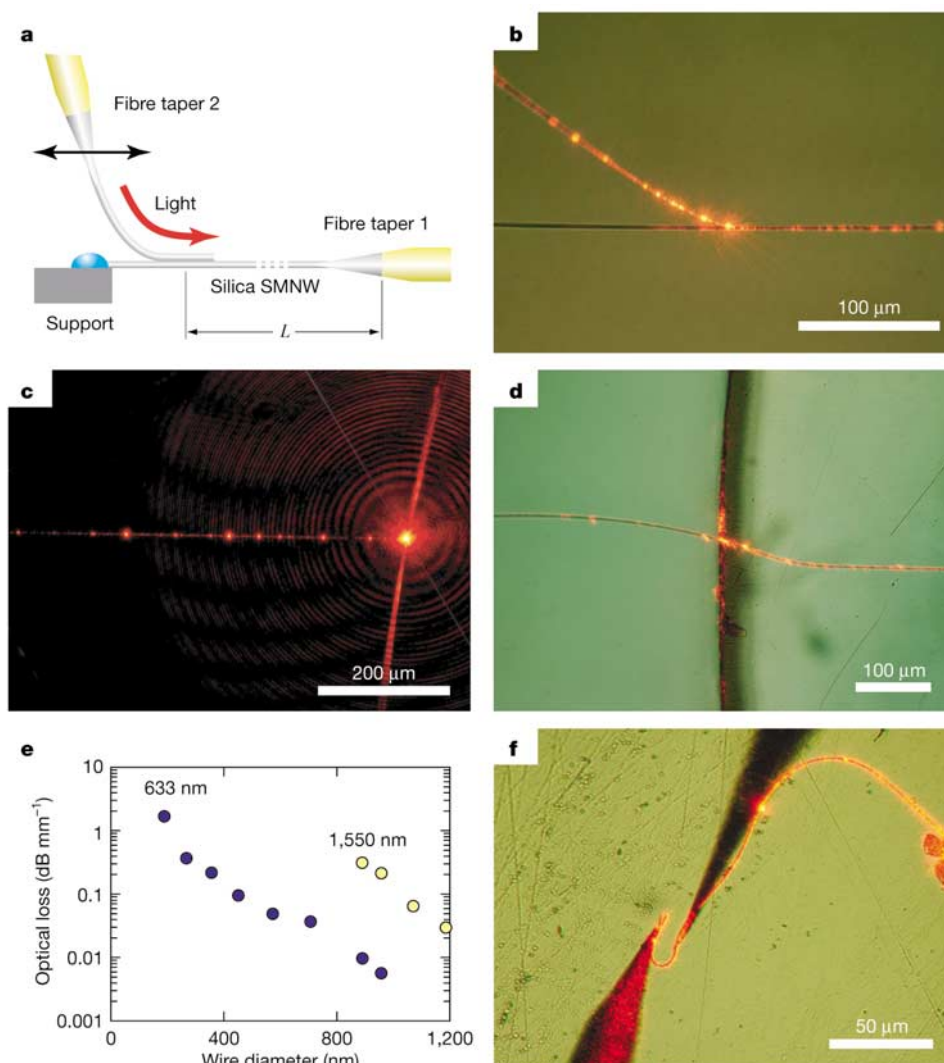


Figure 4 Optical characterization of SMNWs. **a**, Schematic diagram for launching light into a silica wire using evanescent coupling. **b**, Optical microscope image of a 390-nm-diameter taper coupling light into a 450-nm-diameter silica wire. **c**, Long-time exposure micrograph of 633-nm wavelength light guided by a 360-nm-diameter silica wire in air, and intercepted by a $3\text{-}\mu\text{m}$ guiding wire on the right. **d**, Optical microscope image of

633-nm-wavelength light guided by a 550-nm-diameter silica wire with its left half suspended in air and its right half placed on a MgF_2 crystal. **e**, Measured optical loss of silica wires at 633 nm (filled blue circles) and 1,550 nm (filled yellow circles). **f**, Optical microscope image of 633-nm light travelling through a sharp bend with a radius of $5.6 \mu\text{m}$ in a 510-nm-wide silica wire.

The wires reported here are suitable for low-loss optical wave guiding, and will be promising components in future microphonic devices for various applications, such as optical communications and optical sensing. Owing to their excellent uniformity, large length, high flexibility, and strength, these wires can be manipulated and assembled with high accuracy and used as micro- or nanoscale tools in physical, chemical, biological, micro-electronic and materials research. □

Received 22 August; accepted 4 November 2003; doi:10.1038/nature02193.

1. Yamane, M. & Asahara, Y. *Glasses for Photonics* (Cambridge Univ. Press, Cambridge, UK, 2000).
2. Murata, H. *Handbook of Optical Fibers and Cables* 2nd edn (Marcel Dekker, New York, 1996).
3. Mynbaev, D. K. & Scheiner, L. L. *Fiber-Optic Communications Technology* (Prentice Hall, New York, 2001).
4. Marcuse, D. Mode conversion caused by surface imperfections of a dielectric slab waveguide. *Bell Syst. Tech. J.* **48**, 3187–3215 (1969).
5. Marcuse, D. & Derosier, R. M. Mode conversion caused by diameter changes of a round dielectric waveguide. *Bell Syst. Tech. J.* **48**, 3217–3232 (1969).
6. Ladouceur, F. Roughness, inhomogeneity, and integrated optics. *J. Lightwave Technol.* **15**, 1020–1025 (1997).
7. Lee, K. K. *et al.* Effect of size and roughness on light transmission in a Si/SiO₂ waveguide: experiments and model. *Appl. Phys. Lett.* **77**, 1617–1619 (2000). Erratum. *Appl. Phys. Lett.* **77**, 2258 (2000).
8. Boys, C. V. On the production, properties, and some suggested uses of the finest threads. *Phil. Mag.* **23**, 489–499 (1887).
9. Threlfall, R. *On Laboratory Arts* (Macmillan, London, 1898).
10. Knight, J. C., Cheung, G., Jacques, F. & Birks, T. A. Phase-matched excitation of whispering-gallery-mode resonances by a fiber taper. *Opt. Lett.* **22**, 1129–1131 (1997).
11. Birks, T. A., Wadsworth, W. J. & Russell, P. St. J. Supercontinuum generation in tapered fibers. *Opt. Lett.* **25**, 1415–1417 (2000).
12. Cai, M. & Vahala, K. Highly efficient hybrid fiber taper coupled microsphere laser. *Opt. Lett.* **26**, 884–886 (2001).
13. Kakarantzias, G., Dimmick, T. E., Birks, T. A., Le Roux, R. & Russell, P. St. J. Miniature all-fiber devices based on CO₂ laser microstructuring of tapered fibers. *Opt. Lett.* **26**, 1137–1139 (2001).
14. Dimmick, T. E., Kakarantzias, G., Birks, T. A. & Russell, P. St. J. Carbon dioxide laser fabrication of fused-fiber couplers and tapers. *Appl. Opt.* **38**, 6845–6848 (1999).
15. Grellier, A. J. C., Zayer, N. K. & Pannell, C. N. Heat transfer modeling in CO₂ laser processing of optical fibres. *Opt. Commun.* **152**, 324–328 (1998).
16. Wang, Z. L., Gao, R. P. P., Gole, J. L. & Stout, J. D. Silica nanotubes and nanofiber arrays. *Adv. Mater.* **12**, 1938–1940 (2000).
17. Pan, Z. W., Dai, Z. R., Ma, C. & Wang, Z. L. Molten gallium as a catalyst for the large-scale growth of highly aligned silica nanowires. *J. Am. Chem. Soc.* **124**, 1817–1822 (2002).
18. Hu, J. Q., Meng, X. M., Jiang, Y., Lee, C. S. & Lee, S. T. Fabrication of germanium-filled silica nanotubes and aligned silica nanofibers. *Adv. Mater.* **15**, 70–73 (2003).
19. Labelle, H. E. & Mlavsky, A. I. Growth of sapphire filaments from melt. *Nature* **216**, 574–575 (1967).
20. Morales, A. M. & Lieber, C. M. A laser ablation method for the synthesis of crystalline semiconductor nanowires. *Science* **279**, 208–211 (1998).
21. Xia, Y., Rogers, J. A., Paul, K. E. & Whitesides, G. M. Unconventional methods for fabricating and patterning nanostructures. *Chem. Rev.* **99**, 1823–1848 (1999).
22. Ito, T. & Okazaki, S. Pushing the limits of lithography. *Nature* **406**, 1027–1031 (2000).
23. Lee, K. K., Lim, D. R., Kimerling, L. C., Shin, J. & Cerrina, F. Fabrication of ultralow-loss Si/SiO₂ waveguides by roughness reduction. *Opt. Lett.* **26**, 1888–1890 (2001).
24. Matthewson, M. J., Kurkjian, C. R. & Gulati, S. T. Strength measurement of optical fibers by bending. *J. Am. Ceram. Soc.* **69**, 815–821 (1986).
25. Krause, J. T., Testardi, L. R. & Thurston, R. N. Deviations from linearity in the dependence of elongation upon force for fibers of simple glass formers and of glass optical light guides. *Phys. Chem. Glasses* **20**, 135–139 (1979).
26. Takahara, J., Yamagishi, S., Taki, H., Morimoto, A. & Kobayashi, T. Guiding of a one-dimensional optical beam with nanometer diameter. *Opt. Lett.* **22**, 475–477 (1997).
27. Maier, S. A., Kik, P. G. & Atwater, H. A. Observation of coupled plasmon-polarization modes in Au nanoparticle chain waveguides of different lengths: Estimation of waveguide loss. *Appl. Phys. Lett.* **81**, 1714–1716 (2002).
28. Maier, S. A. *et al.* Local detection of electromagnetic energy transport below the diffraction limit in metal nanoparticle plasmon waveguides. *Nature Mater.* **2**, 229–232 (2003).
29. Snyder, A. W. & Love, J. D. *Optical Waveguide Theory* (Chapman and Hall, New York, 1983).
30. Taflov, A. *Computational Electrodynamics: The Finite-difference Time-domain Method* (Artech House, Boston, 1995).

Acknowledgements We thank Y. Lu, Z. Han and B. Tull for assistance in SEM and TEM imaging, and L. Liu and X. Chen for help with numerical simulations. This work was supported by the US National Science Foundation and by the National Natural Science Foundation in China. T.L. acknowledges support from the Centre for Imaging and Mesoscale Structures at Harvard University.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.M. (mazur@deas.harvard.edu).

Synthesis of a Möbius aromatic hydrocarbon

D. Ajami¹, O. Oeckler², A. Simon² & R. Herges¹

¹Institut für Organische Chemie, Universität Kiel, Otto-Hahn-Platz 4, 24098 Kiel, Germany

²Max-Planck-Institut für Festkörperforschung, Heisenbergstraße 1, D-70569 Stuttgart, Germany

The defining feature of aromatic hydrocarbon compounds is a cyclic molecular structure stabilized by the delocalization of π electrons that, according to the Hückel rule, need to total $4n + 2$ ($n = 1, 2, \dots$); cyclic compounds with $4n$ π electrons are antiaromatic and unstable. But in 1964, Heilbronner predicted¹ on purely theoretical grounds that cyclic molecules with the topology of a Möbius band—a ring constructed by joining the ends of a rectangular strip after having given one end half a twist—should be aromatic if they contain $4n$, rather than $4n + 2$, π electrons. The prediction stimulated attempts to synthesize Möbius aromatic hydrocarbons, but twisted cyclic molecules are destabilized by large ring strains, with the twist also suppressing overlap of the p orbitals involved in electron delocalization and stabilization. In larger cyclic molecules, ring strain is less pronounced but the structures are very flexible and flip back to the less-strained Hückel topology^{2,3}. Although transition-state species⁴, an unstable intermediate⁵ and a non-conjugated cyclic molecule⁶, all with a Möbius topology, have been documented, a stable aromatic Möbius system has not yet been realized. Here we report that combining a ‘normal’ aromatic structure (with p orbitals orthogonal to the ring plane) and a ‘belt-like’ aromatic structure (with p orbitals within the ring plane) yields a Möbius compound stabilized by its extended π system.

Numerous theoretical calculations have been performed to predict the properties of potential Möbius aromatic systems^{7–12}. The twist in the π system is usually introduced by a suitable arrangement of E and Z double bonds. *Trans*-benzene¹³ is the smallest conceivable antiaromatic Möbius annulene. It is only a very shallow minimum on the energy hypersurface, with an energy as much as 107 kcal mol^{−1} above benzene¹⁴. The next-higher Möbius homologue, *trans*-cyclooctatetraene, indeed comes energetically closer to its Hückel (all-*cis* and global minimum) isomer¹⁴. Nevertheless, with a relative energy of 21.3 kcal mol^{−1}, it is expected to be extremely difficult to synthesize. Moreover, there is almost no conjugation between the E and the neighbouring Z double bonds

Table 1 Relative stabilities

[16]annulene			Möbius stabilized [16]annulene		
E_{ref}	Topology	Sym.	E_{ref}	Topology	Sym.
0.0	Hückel	S ₄	0.0	Möbius	C ₁
2.0	Hückel	C ₁	0.3	Möbius	C ₂
5.1	Möbius	C ₁	2.8	Möbius	C ₂
7.6	Möbius	C ₂	2.8	Möbius	C ₁
15.8	Möbius	C ₂	6.7	Möbius	C ₁
51.4	Möbius	C ₂	7.0	Hückel	C _s
			8.3	Möbius	C ₁

Calculated relative stabilities (E_{ref} in kcal mol^{−1}) of the most stable isomers of the parent [16]annulene and of our modified [16]annulene at the B3LYP/6-31G* level of DFT.

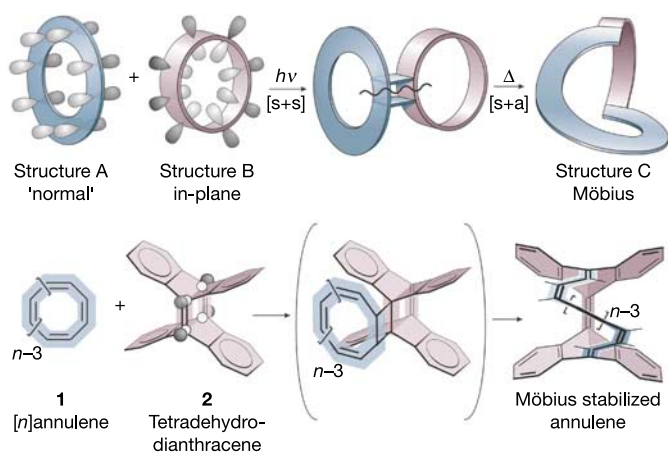


Figure 1 Strategy to stabilize the Möbius structure of annulenes. Möbius bands (structure C) contain a 'normal' trigonal planar conjugated part (shown cyan) and a pyramidalized 'in-plane' part (magenta). The strain arises from the latter. Combining a normal conjugated ring (structure A; for example, $[n]$ annulene **1**) with a rigid 'prefabricated' pyramidalized in-plane conjugated building block (structure B; for example, tetrahydrodianthracene **2**) by metathesis leads to rings that are more stable in the twisted Möbius configuration than in the untwisted Hückel configuration. A supra supra allowed $[2 + 2]$ addition ($[s + s]$) followed by a thermal supra antara cycloreversion ($[s + a]$) leads to a Möbius system.

because of the large deformation of the ring induced by the twist. Castro and Schleyer¹⁵ have published an extensive study of various isomers and conformations of higher annulenes. One of the most interesting structures is a [16]annulene with an $E,Z,Z,Z,Z,Z,E,E,Z,E,E,Z,Z,Z,Z$ arrangement of conjugated bonds that exhibits distinct Möbius aromaticity. The relative energy is $15.3 \text{ kcal mol}^{-1}$ above the global minimum isomer. In view of the large number of conceivable isomers—many of which are more stable than this strongly conjugated Möbius compound—the directed synthesis or isolation of this compound is expected to be a very difficult task.

Our synthetic approach is therefore based on a general strategy that stabilizes Möbius structures (Fig. 1). In 'normal' aromatic structures like benzene, the p orbitals are orthogonal with respect to the ring plane (Fig. 1, structure A, cyan) whereas in belt-like systems (for example, carbon nanotubes) the p orbitals are perpendicular with respect to the surface of a cylinder (Fig. 1, structure B, magenta). A Möbius system (Fig. 1, structure C) includes both types of conjugation: a 'normal' part with trigonal planar sp^2 -hybridized atoms, and a belt-like system with pyramidalized sp^2 atoms. The strain arises from pyramidalization. Thus, introducing a 'prefabricated' pyramidalized building block (Fig. 1, bottom, magenta) in a ring should stabilize the Möbius- versus Hückel-type isomers.

We chose bianthraquinodimethane as this rigid in-plane conjugated part. The steric hindrance of the inner *ortho* protons of the anthracene units enforces the pyramidalization of the central quinoid bond. Ring enlargement metathesis of tetrahydrodianthracene (TDDA) **2** (refs 16–18) with cyclooctatetraene **1** ($n = 1$) should lead to the twisted annulenes.

To test our strategy, preliminary theoretical calculations were performed on the Möbius stabilized [16]annulene. The geometrically fixed anthraquinodimethane part reduces the number of conceivable E/Z isomers to 12, each of which can have 9 s -*cis*/ s -*trans* conformations. We optimized all 108 isomers using the PM3 method¹⁹, and performed density functional theory (DFT) calculations²⁰ at the B3LYP/6-31G* level^{21,22} of theory on the most stable conformation of the 12 most stable isomers. In Table 1 the relative energies of our Möbius stabilized annulene are compared with the results of Castro and Schleyer¹⁵ on the parent [16]annulene.

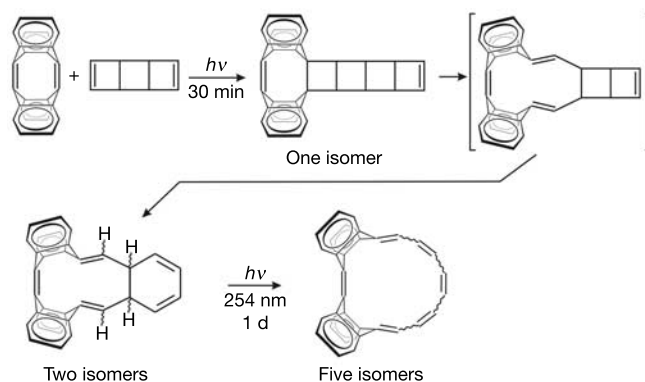


Figure 2 Reaction of *syn*-tricyclocyclooctadiene with tetrahydrodianthracene.

Whereas in the parent [16]annulene the two most stable isomers have Hückel topology, the six most stable isomers of our modified [16]annulene are Möbius-type systems including the global minimum. Hence, the theoretical data confirm our Möbius stabilizing strategy, and on the basis of these thermodynamic data, the synthesis of a Möbius system should be feasible.

Unfortunately, all attempts to combine TDDA with cyclooctatetraene in a ring enlargement metathesis reaction failed. Even though benzene and a number of alkenes upon irradiation react with TDDA to form metathesis products, cyclooctatetraene under the same conditions only furnished Diels–Alder adducts, isomers of cyclooctatetraene and 9,9'-bianthracene. To circumvent these problems, we used *syn*-tricyclocyclooctadiene (TCOD) as a synthetic equivalent of cyclooctatetraene²³. Irradiation of TDDA and TCOD in benzene with a high-pressure mercury lamp, using a quartz apparatus, gave as the main products two 1,3-cyclohexadiene derivatives (Fig. 2; see Supplementary Information for a detailed experimental procedure).

In the first step, a $[2 + 2]$ cycloaddition product is formed, which could be isolated and characterized by NMR. This ladder-shaped compound, containing four fused cyclobutane rings, undergoes cycloreversion to give an unstable intermediate, which again opens in an electrocyclic reaction to form a C_2 and a C_s symmetric 1,3-cyclohexadiene. Both isomers were isolated, and characterized by X-ray structure analysis. Only minor amounts of the desired fully ring opened products were detected. As it is known that the photochemical stationary equilibrium between 1,3-cyclohexadienes and hexatrienes is shifted towards the ring opened product at shorter wavelengths, we irradiated the reaction mixture for 24 h with a low-pressure mercury lamp and obtained a mixture of isomers in an overall combined yield of about 50%. Five ring opened isomers were isolated by high-performance liquid chromatography. Two of them exhibit Möbius topology (C_1 and C_2 symmetry) and one has Hückel topology (C_s symmetry). The two Möbius structures were elucidated by NMR, and confirmed by X-ray analysis (for details of the X-ray data of the C_2 Möbius compound, see Supplementary Information; Cambridge Crystallographic Data Centre deposition number 223927). Both geometries are in very good agreement with DFT calculations. Only the C_2 symmetry isomer is significantly conjugated. The structural parameters of the Hückel C_s structure are not accurate because of disorder ($<10\%$ of a second isomer), but the data confirm the structural assignment by NMR. We use the Hückel structure as a reference to investigate the properties of our new Möbius C_2 compound (Fig. 3).

The C_2 symmetry Möbius compound forms red crystals, whereas the C_s symmetry Hückel compound is colourless (for ultraviolet data, see Supplementary Information). In Fig. 3, the most important geometry parameters calculated at the B3LYP/6-31G* level of

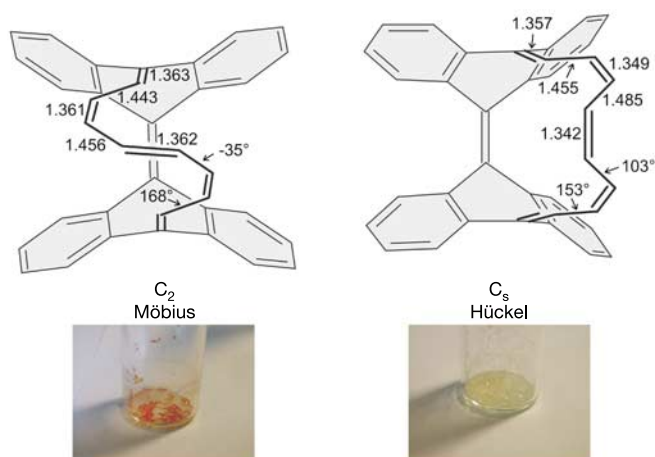


Figure 3 X-ray structures and photographs of the crystals of the C_2 Möbius and the C_s Hückel isomer. The structure plots are idealized for clarity (thermal ellipsoids are removed, and the bond orders are indicated by single and double lines). Structural parameters for the Hückel isomer are not reliable because of disorder (see text). Therefore the bond lengths (Å) and $C=C-C=C$ dihedral angles for the polyene bridge shown in the figure are determined by DFT calculations (B3LYP/6-31G*). For the complete set of X-ray data and calculations, see Supplementary Information.

DFT are given. Whereas the rigid bianthraquinodimethane part is almost identical in both isomers, the geometry of the polyene bridge differs substantially. Bond-length equalization, which is used as an aromaticity probe, is more pronounced in the Möbius compound than in the Hückel structure. The HOMA value²⁴, which quantifies this effect, is 0.50 in the polyene bridge (9 bonds) of the Möbius isomer, and 0.05 in the Hückel isomer. For comparison, the prototype aromatic benzene has a HOMA value of 0.98. The total HOMA (all bonds included) of the Möbius isomer is 0.35, and the HOMA of the Hückel structure is 0.17. A value of 0.35 corresponds to the HOMA in the benzene rings of C_{60} , and is somewhat lower than the corresponding value of the central benzene ring in phenanthrene (0.40). The Hückel isomer escapes from the 16-electron antiaromaticity by twisting the central double bond in the polyene bridge by 103° with respect to the two neighbouring double bonds. The dihedral angles in the Möbius system, however, only differ 12° and 35° from the ideal *s-cis* or *s-trans* conformation, still allowing considerable conjugation.

Aromaticity is a multidimensional property²⁵, and arguments exclusively based on geometric parameters may be misleading. For example, the maximum bond-length difference between single and double bonds in our C_2 Möbius compound (0.095 Å) is still quite large, comparable to the value for the non-aromatic anti-bismethano-[14]annulene²⁶. We therefore also investigated a second, independent, aromaticity measure based on energy considerations, by using the 'indene-isoidene method' (ref. 27) to estimate the stabilization caused by conjugation in the Möbius structure. At the B3LYP/6-31G* level of DFT we computed a stabilization energy due to conjugation, ISE_{II} , of $4.04 \text{ kcal mol}^{-1}$ for the C_2 Möbius compound (20% of the ISE_{II} of benzene) and an antiaromatic destabilization energy of $-2.22 \text{ kcal mol}^{-1}$ for the Hückel C_s structure. Both values are uncorrected, and can be expected to be more positive if *syn-anti* corrections are included²⁸. We also find that the strain energy of the Möbius compound is slightly higher than that of the Hückel isomer, clearly indicating that it is the more efficient conjugation that makes the Möbius compound more stable than the Hückel compound.

Taken together, the observed trends in bond-length equalization and stabilization energy indicate that the Möbius structure is moderately aromatic, whereas the Hückel structure is non-

aromatic. In the C_2 Möbius structure, all C-C bonds are conjugated, yet the molecule has only one π surface. The properties investigated here finally confirm Heilbronner's prediction for Möbius twisted $[4n]$ annulenes. □

Received 18 September; accepted 21 November 2003; doi:10.1038/nature02224.

- Heilbronner, E. Hückel molecular orbitals of Möbius-type conformations of annulenes. *Tetrahedr. Lett.* **29**, 1923–1928 (1964).
- Garrat, P. J. *Aromaticity* (Wiley and Sons, New York, 1986).
- Oth, J. F. M. & Gilles, J.-M. Mobilité conformationnelle et isomère de valence rapide réversible dans le [16]annulène. *Tetrahedr. Lett.* **60**, 6259–6264 (1968).
- Zimmerman, H. E. On molecular orbital correlation diagrams, the occurrence of Möbius systems in cyclization reactions. *J. Am. Chem. Soc.* **88**, 1564–1567 (1966).
- Mauksch, M., Gogonea, V., Jiao, H. & von Ragué Schleyer, P. Monocyclic $(CH)_8^+$ —a Heilbronner-Möbius arene. *Angew. Chem. Int. Edn Engl.* **37**, 2395–2398 (1998).
- Walba, D. M., Homan, T. C., Richards, R. M. & Haltiwanger, R. C. Topological stereochemistry 9. Synthesis and cutting "in half" of a molecular Möbius strip. *New J. Chem.* **17**, 661–681 (1993).
- Dobrowolski, J. Cz. On the belt and Möbius isomers of the coronene molecule. *J. Chem. Inf. Comput. Sci.* **42**, 490–499 (2002).
- Guillaume, M., Champagne, B., Perpète, E. A. & André, J.-M. Möbius strip versus linear and cyclic polyacenes: a Hückel and semiempirical investigation. *Theor. Chem. Acc.* **105**, 431–436 (2001).
- Zoellner, R. W., Krebs, J. F. & Browne, D. M. Violently twisted and strained organic molecules. *J. Chem. Inf. Comput. Sci.* **34**, 252–258 (1994).
- Kastrup, C. J., Oldfield, S. V. & Rzepa, H. S. An ab initio study of monodentate palladium ligand complexes with Möbius-aromatic chiral character. *J. Chem. Soc. Dalton Trans.* 2421–2422 (2002).
- Martin-Santamaría, S., Lavan, B. & Rzepa, H. S. Möbius aromatics arising from a $C=C=C$ ring component. *Chem. Commun.* 1089–1090 (2000).
- Martin-Santamaría, S. & Rzepa, H. S. Twist localization in single, double and triple twisted Möbius cyclozenes. *J. Chem. Soc. Perkin Trans. 2*, 1415–1417 (2000).
- Johnson, R. P. & Daoust, K. J. Electrocyclic ring opening modes of Dewar benzenes: ab initio predictions for Möbius benzene and *trans*-Dewar benzene as new C_6H_6 isomers. *J. Am. Chem. Soc.* **118**, 7381–7385 (1996).
- Havenith, R. W. A., van Lenthe, J. H. & Jenneskens, L. W. Möbius Aromaticity in small $[n]$ trans annulenes? *Int. J. Quant. Chem.* **85**, 52–60 (2001).
- Castro, C., Isborn, C. M., Karney, W. L., Mauksch, M. & von Ragué Schleyer, P. Aromaticity with a twist: Möbius $[4n]$ annulenes. *Org. Lett.* **4**, 3431–3434 (2002).
- Kammermeier, S. & Herges, R. Photochemically induced metathesis reactions of tetrahydrodianthracene: synthesis and structure of bianthraquinodimethanes. *Angew. Chem.* **108**, 470–472 (1996); *Angew. Chem. Int. Edn Engl.* **35**, 417–419 (1996).
- Kammermeier, S., Jones, P. G. & Herges, R. Ring-expanding metathesis of tetrahydrodianthracene—synthesis and structure of a tubelike, fully conjugated hydrocarbon. *Angew. Chem.* **108**, 2834–2836 (1996); *Angew. Chem. Int. Edn Engl.* **35**, 2669–2671 (1996).
- Kammermeier, S. & Herges, R. Beltlike aromatic hydrocarbons by metathesis reaction with tetrahydrodianthracene. *Angew. Chem.* **109**, 2317–2319 (1997); *Angew. Chem. Int. Edn Engl.* **36**, 2200–2202 (1997).
- Stewart, J. J. P. Optimization of parameters for semi-empirical methods I. Method. *J. Comput. Chem.* **10**, 209–220 (1989).
- Frisch, M. J. *et al.* Gaussian 98, Revision A.11.4 (Gaussian, Inc., Pittsburgh, PA, 2002).
- Becke, A. D. Density-functional thermochemistry. III. The role of exact exchange. *J. Chem. Phys.* **98**, 5648–5652 (1993).
- Lee, C., Yang, W. & Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **37**, 785–789 (1988).
- Avram, M. *et al.* Zwei stereoisomere Dimere des Cyclobutadiens. *Chem. Ber.* **97**, 382–389 (1964).
- Krygowski, T. M. Crystallographic studies of inter- and intramolecular interactions reflected in aromatic character of π electron systems. *J. Chem. Inf. Comput. Sci.* **33**, 70–78 (1993).
- Katritzky, A. R., Jug, K. & Oniciu, C. D. Quantitative measures of aromaticity for mono-, bi- and tricyclic penta- and hexatomic heteroaromatic ring systems and their interrelationships. *Chem. Rev.* **101**, 1421–1449 (2001).
- Nendel, M. R. *et al.* Bond alternation and aromatic character in cyclic polyenes: assessment of theoretical methods for computing the structures and energies of bismethano[14]annulenes. *J. Phys. Chem. A* **102**, 7191–7198 (1998).
- Wannere, C. S. *et al.* On the stability of large $[4n]$ annulenes. *Org. Lett.* **5**, 2983–2986 (2003).
- von Schleyer, P. R. & Pühlhofer, F. Recommendations for the evaluation of aromatic stabilization energies. *Org. Lett.* **4**, 865–868 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank C. Näther for measuring the X-ray structures of the Hückel C_s and the Möbius C_2 compounds, and C. Wolff for the interpretation of the NMR spectra. We also thank E. Heilbronner for providing additional information on ref. 1, and T. Bally for hints on how to improve the synthesis of *syn*-tricyclocloctadiene. D.A. was supported by a scholarship from the federal state of Niedersachsen.

Authors' contributions R.H. together with D.A. conceived the experiment; D.A. carried it out; R.H. wrote the Letter; and A.S. and O.O. determined the X-ray structure of the Möbius C_2 compound.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.H. (rherges@oc.uni-kiel.de).

Stable isotopic evidence for methane seeps in Neoproterozoic postglacial cap carbonates

Ganqing Jiang¹, Martin J. Kennedy¹ & Nicholas Christie-Blick²

¹Department of Earth Sciences, University of California, Riverside, California 92521, USA

²Department of Earth and Environmental Sciences and Lamont-Doherty Earth Observatory of Columbia University, Palisades, New York 10964-8000, USA

The Earth's most severe glaciations are thought to have occurred about 600 million years ago, in the late Neoproterozoic era^{1,2}. A puzzling feature of glacial deposits from this interval is that they are overlain by 1–5-m-thick 'cap carbonates' (particulate deep-water marine carbonate rocks) associated with a prominent negative carbon isotope excursion^{3–8}. Cap carbonates have been controversially ascribed to the aftermath of almost complete shutdown of the ocean ecosystems for millions of years during such ice ages—the 'snowball Earth' hypothesis^{4,5}. Conversely, it has also been suggested that these carbonate rocks were the result of destabilization of methane hydrates during deglaciation and concomitant flooding of continental shelves and interior basins³. The most compelling criticism of the latter 'methane hydrate' hypothesis has been the apparent lack of extreme isotopic variation in cap carbonates inferred locally to be associated with methane seeps. Here we report carbon isotopic and petrographic data from a Neoproterozoic postglacial cap carbonate in south China that provide direct evidence for methane-influenced

processes during deglaciation. This evidence lends strong support to the hypothesis that methane hydrate destabilization contributed to the enigmatic cap carbonate deposition and strongly negative carbon isotopic anomalies following Neoproterozoic ice ages. This explanation requires less extreme environmental disturbance than that implied by the snowball Earth hypothesis^{4,5}.

Methane oxidized in sea-floor sediments is hypothesized to have provided both a source for more than 10¹⁷ mol of excess CO₃^{2–} driving carbonate precipitation and a plausible explanation for the timing (<10⁵ yr) and magnitude of the observed carbon isotope anomaly³. This hypothesis gains support from a suite of distinctive sedimentary structures that are found locally within many cap carbonates, and are known collectively only in seep environments^{3,6}.

Carbonate precipitated in modern seeps and in at least some ancient seeps is typically associated with δ¹³C values of less than –25‰, and in some cases as low as –50‰ (refs 9–12). Heavier δ¹³C values ranging from –24‰ to +6‰ are also common in many examples^{9,10,12}, depending on the degree of ambient seawater involvement in carbonate precipitation, and as a result of methanogenesis during burial^{13,14}. In some cases, pervasive diagenesis results in homogenization of isotopic values, thereby obscuring the isotopic contrast between the host rock precipitated from sea water and seep-related carbonate cements. In Miocene seeps exposed near Santa Cruz, California, δ¹³C values for most of the calcite samples analysed are between –4‰ to –10‰ (ref. 15). Such homogenization at a scale of millimetres to centimetres is expected for cap carbonates, most of which are now composed of dolomite. The absence of evidence for extreme isotopic variation in any cap carbonate, for which there is now a substantial global data set^{3–5,6–8,16}, has nevertheless been viewed as a serious shortcoming.

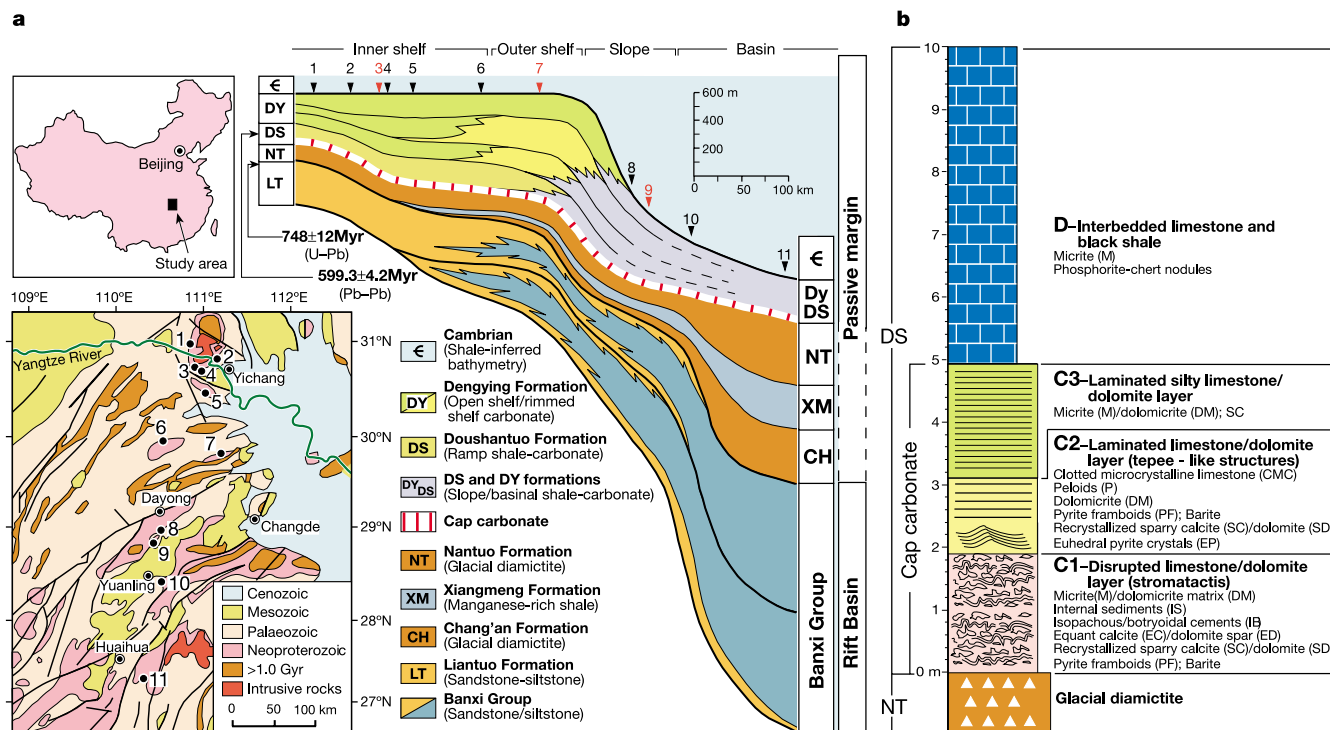


Figure 1 Stratigraphy and cap carbonate microfacies. **a**, Stratigraphic cross-section of the Yangtze platform in south China; **b**, summary of microfacies in the cap carbonate. Section numbers matching those in Figs 3 and 4 are in red. Note that the cap carbonate

and overlying deposits are preferentially dolomitized in inferred deeper-water sections (Fig. 4).

To determine the degree to which diagenesis has obscured distinctive isotopic signals imparted by methane-related processes, we conducted a detailed isotopic and petrographic study of seep-like features in the cap carbonate that overlies the glaciogenic Nantuo Formation in the Yangtze platform of south China (Fig. 1a). Although this cap carbonate is in many respects typical of such carbonates globally, it is notable for its unusually low thermal maturity, degree of textural preservation, localized retention of calcite mineralogy, and abundance of well-exposed sections in a platform to basin transect. The Nantuo Formation is the younger of two Neoproterozoic glacial diamictite units preserved within a rift to passive margin succession (Fig. 1a), and is inferred to have an age greater than ~600 Myr (refs 17, 18). It underlies shale and limestone of the Doushantuo Formation, in which some of the earliest animal fossils have been identified¹⁸.

Located at the base of the Doushantuo, the 3–5-m-thick cap carbonate contains three lithologically distinguishable but laterally variable limestone and dolomite units (Fig. 1b): a basal layer, 1.0–1.9 m thick, which is strongly disrupted and cemented (C1), a middle, laminated layer, less than 1 m thick, with local ‘tepee-like’ structures (C2), and an upper layer, 1.5–2 m thick, composed of thinly laminated silty limestone and dolomite (C3). Irregular cavities formed by subhorizontal extension of micritic crusts, and filled with multiple generations of fringing (commonly botryoidal) cements, resemble features referred to as stromatactis in some ancient methane-seep deposits^{10,12}. Structures described here as tepee-like are of positive relief and from 1–3 m wide and <1 m high, composed of laminated micrite and dolomicrite with layer-parallel sheet cracks, and cored by broken and brecciated host carbonate. Voids in these structures are lined by fringing botryoidal and clotted cements with pyrite framboids (Fig. 2), including radiating blade-shaped crystals of barite (<5 mm; mostly replaced by calcite), and they are filled at least partially by internal sediments. We infer on this basis that the structures are syn-sedimentary, and that the voids remained open to the sea floor during development.

Isotopic analysis of limestone peloids, clots and fringing cements within and above the tepee-like structures where they are best preserved (location 3 in Fig. 1) reveals highly variable $\delta^{13}\text{C}$ values, ranging from +5‰ to –41‰ (Fig. 3). These data represent, to our knowledge, the first direct geochemical support for deposition in the vicinity of a methane seep. Carbon isotopic values as low as –41‰ in carbonate carbon are known only from carbonate precipitation associated with the oxidation of methane, the carbon isotopic composition of which is the lightest known in nature¹¹. The documented variability of carbon isotopic values is characteristic of seep facies^{9–12}. The strong inverse correlation with $\delta^{18}\text{O}$ ($R^2 = 0.93$, $n = 16$; Fig. 3) indicates that the most negative $\delta^{13}\text{C}$ values, at least, approach a primary value^{13,14}.

Isotopically heavy $\delta^{13}\text{C}$ (+0.95 to +4.98‰) and $\delta^{18}\text{O}$ (–1.1 to +3.0‰) values (interval C2) and abrupt excursions (each defined by several data points) in $\delta^{13}\text{C}$ (–4.7 to +4.7‰) and $\delta^{18}\text{O}$ (–12.2 to –1.5‰) (interval C3) above the tepee-like structures reinforce the link to methane-related phenomena. They are consistent with either microbially mediated methane metabolism^{13,14} or with the decomposition of gas hydrates during burial diagenesis^{10,13,14}. Methane would have been partially metabolized by bacterial sulphate reduction to produce monosulphides and alkalinity¹⁹, a process that is consistent with the presence of framboidal pyrite in cements (Fig. 2a and c). The presence of barite reinforces similarities with other cap carbonate deposits^{3,5} and modern seeps²⁰, and has been linked to massive dissociation of methane hydrates at the Late Palaeocene Thermal Maximum (LPTM)²¹.

While this cap carbonate provides clear if local evidence for methane seeps, it also opens a window into the modification of the isotopic signature of cap carbonates through diagenesis. Carbon and oxygen isotope values from interval C2 in Fig. 3 fall on a line

between a methane-influenced end member (–41‰ $\delta^{13}\text{C}$; –4.3‰ $\delta^{18}\text{O}$) and spar-filling tectonic fractures (–3.5‰ $\delta^{13}\text{C}$; –13.5‰ $\delta^{18}\text{O}$). Mixing along this line implies partial alteration of the highly negative values of the seep facies towards a regional diagenetic fluid value preserved in late-stage cross-cutting veins. More pervasively dolomitized deeper-water sections (Fig. 4) display none of the

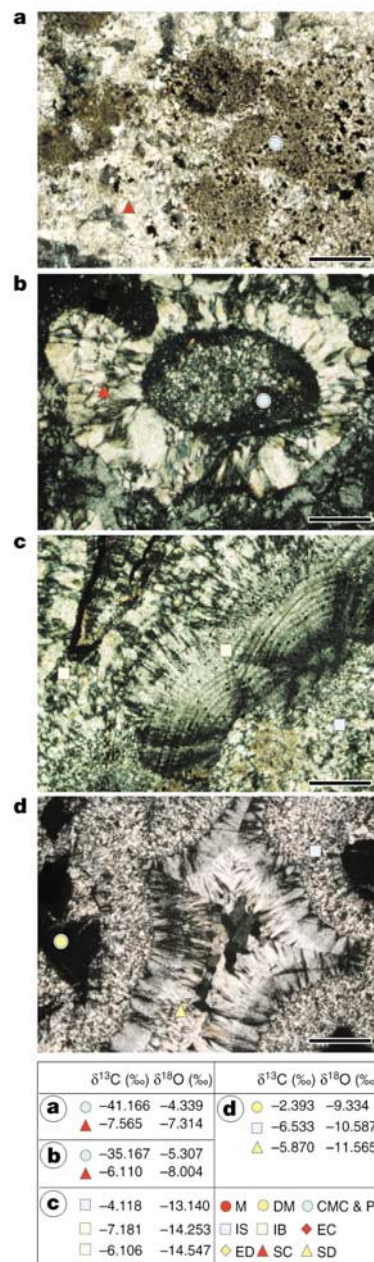


Figure 2 Cap carbonate texture and isotopic composition. **a**, Clots surrounded by sparry calcite containing framboidal pyrite (small black spots). **b**, Peloids surrounded by recrystallized calcite spar. **c**, Partially recrystallized internal sediment and botryoidal cement with radial aragonite fabric now replaced by dolomite. Small black spots within cement are pyrite framboids. **d**, Brecciated dolomicrite matrix and microcrystalline internal sediment partially filling cavity. Remaining space is filled by sparry dolomite. **a** and **b** are from interval C2 in Fig. 3; **c** and **d** are from intervals C1 and C2 in Fig. 4b. Isotopic values of microsamples are listed in table at bottom. Symbols indicate sample location but not sample size. Scale bar, 400 μm . Plane-polarized light. See Fig. 1b for microfacies abbreviations.

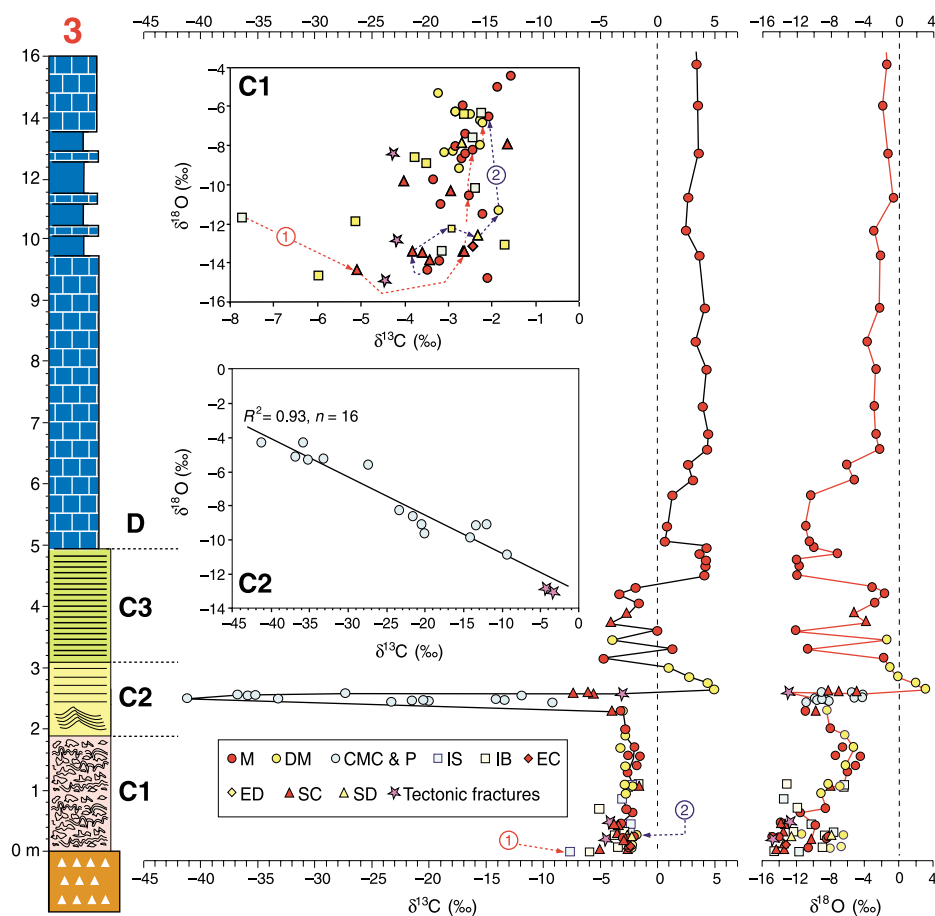


Figure 3 Isotopic composition of cap carbonate from section 3 in Fig. 1a. Carbon isotopic values ranging from -41‰ to $+5\text{‰}$ in C2 are inferred to indicate methane seeps. Connected points in $\delta^{13}\text{C}$ – $\delta^{18}\text{O}$ crossplot for interval C1 are microsamples from individual

polished slabs ($< 10 \times 10 \text{ cm}^2$), stratigraphically located as indicated. See Fig. 1b for microfacies abbreviations.

extreme isotopic variability seen in Fig. 3, even though the same distinct fringing botryoidal, pyrite-bearing cements and tepee-like structures are present also in those sections. The 1–3‰ variation in $\delta^{13}\text{C}$ between sections and covariance between $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$ ($R^2 = 0.81$, $n = 26$; Fig. 4a), and variability of up to 3–5‰ between matrix and cements at millimetre (Fig. 2c and d) and centimetre scales (Figs 3 and 4) suggest substantial diagenetic modification of the isotopic signature for the majority of samples, possibly through dolomitization. This raises questions about the retention of primary isotopic values of dolomitized cap carbonates.

Our data indicate a spatial association between methane-seep sedimentary features and the unusually large global isotopic excursion (-4‰ to -7‰ shift in $\delta^{13}\text{C}$)^{3,5,7,8}. These seep-like facies are preserved only locally within cap carbonates, consistent with either isolated sources or discrete pathways for methane. However, such facies are also widespread and, in all of the examples known to us, the distinctive structures and textures are localized stratigraphically within the lower portion of the cap carbonate. This precludes the interpretation of these seeps as surface expressions of long-lived hydrocarbon systems. In the light of our new data, the most plausible explanation for the broad but stratigraphically restricted distribution of interpreted methane seeps and the carbon isotopic signature of cap carbonates in general is that they both relate to the postglacial destabilization of gas hydrates.

Mass-balance considerations²² for a -4‰ to -7‰ $\delta^{13}\text{C}$

excursion imply the addition of $4.0 \times 10^{18} \text{ g}$ to $7.4 \times 10^{18} \text{ g}$ of ^{13}C -depleted CH_4 to the exchangeable carbon reservoir. This is equivalent to between $\sim 28\%$ and $\sim 53\%$ of the estimated modern gas-hydrate pool. Given estimates of changes in that pool of 14% for the LPTM^{22,23} and 14–24% for the Jurassic Oceanic Anoxic Event²⁴, late Neoproterozoic cap carbonates are inferred to represent the largest hydrate-dissociation event in Earth's history.

An unresolved issue concerns the relative roles of two discrete components of the gas-hydrate inventory³. Although two orders of magnitude smaller than the marine hydrate pool today²⁵, the terrestrial permafrost pool would have been substantially larger during times of extreme Neoproterozoic glaciation, and less sensitive than marine hydrates to the trade-off between increasing temperature and rising pressure (water depth), which may stabilize marine hydrates during times of postglacial sea-level rise. However, inferred seep-related structures and textures are present in the cap carbonate of south China as far south as Huaihua (location 11 in Fig. 1a), where facies and a greatly expanded thickness of underlying glaciogenic diamictite ($> 500 \text{ m}$) suggest accumulation in a marine basin that was never subaerially exposed. Similar features are also observed in slope facies of the Congo craton in Namibia^{3,26}. This raises the possibility that postglacial warming was sufficiently great even in the deep ocean to destabilize the entire gas-hydrate inventory. If methane hydrates play a role in the global carbon cycle

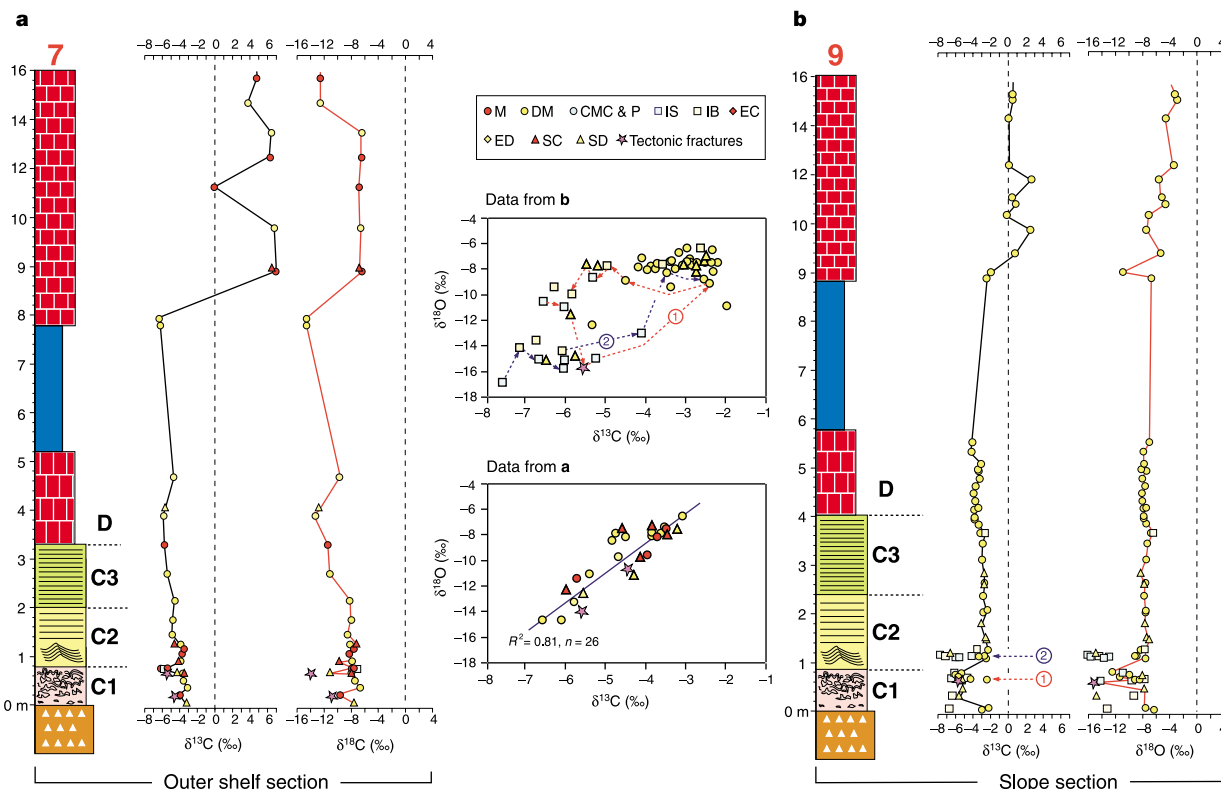


Figure 4 Isotopic composition of cap carbonate from sections 7 and 9 in Fig. 1a. **a**, Section 7; **b**, section 9. Carbon isotopic values are less variable than in Fig. 3. A systematic difference in $\delta^{13}\text{C}$ of 1–3‰ between the sections, from interval C2 to lower part of interval D, suggests diagenesis. Connected points in the $\delta^{13}\text{C}$ – $\delta^{18}\text{O}$ crossplot

(interval C1 to lower D) for section 9 are microsamples from polished slabs ($<10 \times 10 \text{ cm}^2$), stratigraphically located as indicated. See Fig. 1b for microfacies abbreviations.

and climate change, as increasingly appears to be the case, Neoproterozoic geology offers a unique opportunity for investigating their influence, along with an important new aspect of the physical context in which metazoans first appear in the fossil record. □

Methods

Petrographic microscopy, cathodoluminescence (CL), and scanning electron microscopy (SEM) were used for petrographic characterization. Powders (0.5–5 mg) were obtained from polished slabs using 0.5–1.2-mm-diameter microdrills after petrographic study. Sample powders were reacted with 100% phosphoric acid at 50 °C for 2 h. $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$ were determined at the University of California, Riverside, and are reported using the standard δ -notation compared to PDB. Precision of an internal standard for both $\delta^{13}\text{C}$ and $\delta^{18}\text{O}$ was $\pm 0.2\text{‰}$ ($n = 25$). Key samples were reanalysed at the University of California, Davis, as a check of reproducibility. Results from the two laboratories agree within 0.5‰ for $\delta^{13}\text{C}$ and 0.7‰ for $\delta^{18}\text{O}$.

Received 22 June; accepted 10 November 2003; doi:10.1038/nature02201.

1. Crowell, J. C. Pre-Mesozoic ice ages: their bearing on understanding the climate system. *Mem. Geol. Soc. Am.* **192** (1999).
2. Sohl, L. E., Christie-Blick, N. & Kent, D. V. Paleomagnetic polarity reversals in Marinoan (ca 600 Ma) glacial deposits of Australia: implications for the duration of low-latitude glaciations in Neoproterozoic time. *Geol. Soc. Am. Bull.* **111**, 1120–1139 (1999).
3. Kennedy, M. J., Christie-Blick, N. & Sohl, L. E. Are Proterozoic cap carbonates and isotopic excursions a record of gas hydrate destabilization following Earth's coldest intervals? *Geology* **29**, 443–446 (2001).
4. Hoffman, P. F., Kaufman, A. J., Halverson, G. P. & Schrag, D. P. A Neoproterozoic snowball Earth. *Science* **281**, 1342–1346 (1998).
5. Hoffman, P. F. & Schrag, D. P. The snowball Earth hypothesis: testing the limits of global change. *Terra Nova* **14**, 129–155 (2002).
6. Kennedy, M. J. Stratigraphy, sedimentology, and isotopic geochemistry of Australian Neoproterozoic postglacial cap dolostones: deglaciation, $\delta^{13}\text{C}$ excursions, and carbonate precipitation. *J. Sedim. Res.* **66**, 1050–1064 (1996).
7. Kennedy, M. J., Christie-Blick, N. & Prave, A. R. Carbon isotopic composition of Neoproterozoic glacial carbonates as a test of paleoceanographic models for snowball Earth phenomena. *Geology* **29**, 1135–1138 (2001).

8. James, N. P., Narbonne, G. M. & Kyser, T. K. Late Neoproterozoic cap carbonates: Mackenzie Mountains, northwestern Canada: precipitation and global glacial meltdown. *Can. J. Earth Sci.* **38**, 1229–1262 (2001).
9. Stakes, D. S., Orange, D., Paduan, J. B., Salamy, K. A. & Maher, N. Cold-seeps and authigenic carbonate formation in Monterey Bay, California. *Mar. Geol.* **159**, 93–109 (1999).
10. Peckmann, J., Goedert, J. L., Thiel, V., Michaelis, W. & Reitner, J. A comprehensive approach to the study of methane-seep deposits from the Lincoln Creek Formation, western Washington State, USA. *Sedimentology* **49**, 855–873 (2002).
11. Cavagna, S., Clari, P. & Martire, L. The role of bacteria in the formation of cold seep carbonates: geological evidence from Monferrato (Tertiary, NW Italy). *Sedim. Geol.* **126**, 253–270 (1999).
12. Kauffman, E. G., Arthur, M. A., Howe, B. & Scholle, P. A. Widespread venting of methane-rich fluids in Late Cretaceous (Campanian) submarine springs (Tepee Buttes), Western Interior seaway, U.S.A. *Geology* **24**, 799–802 (1996).
13. Irwin, H., Curtis, C. & Coleman, M. Isotopic evidence for source of diagenetic carbonates formed during burial of organic-rich sediments. *Nature* **269**, 209–213 (1977).
14. Gautier, D. L. & Claypool, G. E. Interpretation of methanogenic diagenesis in ancient sediments by analogy with processes in modern diagenetic environments. *Mem. Am. Assoc. Petrol. Geol.* **37**, 111–123 (1984).
15. Aiello, I. W., Garrison, R. E., Moore, J. C., Kastner, M. & Stakes, D. S. Anatomy and origin of carbonate structures in a Miocene cold-seep field. *Geology* **29**, 1111–1114 (2001).
16. Grotzinger, J. P. & Knoll, A. H. Anomalous carbonate precipitates: is the Precambrian the key to the Permian? *Palaios* **10**, 578–596 (1995).
17. Barfod, G. H. & Baker, J. New Lu–Hf and Pb–Pb age constraints on the earliest animal fossils. *Earth Planet. Sci. Lett.* **201**, 203–212 (2002).
18. Xiao, S., Zhang, Y. & Knoll, A. H. Algae and embryos in a Neoproterozoic phosphorite. *Nature* **391**, 553–558 (1998).
19. Boettius, A. et al. A marine microbial consortium apparently mediating anaerobic oxidation of methane. *Nature* **407**, 623–626 (2000).
20. Greinert, J., Bollwerk, S. M., Derkachev, A., Bohrmann, G. & Suess, E. Massive barite deposits and carbonate mineralization in the Derugin Basin, Sea of Okhotsk: precipitation processes at cold seep sites. *Earth Planet. Sci. Lett.* **203**, 165–180 (2002).
21. Dickens, G. R., Fewless, T., Thomas, E. & Bralower, T. J. Excess barite accumulation during the Paleocene-Eocene Thermal Maximum: Massive input of dissolved barium from seafloor gas hydrate reservoirs. *Spec. Pap. Geol. Soc. Am.* **369**, 11–23 (2003).
22. Dickens, G. R., O'Neil, J. R., Rea, D. K. & Owen, R. M. Dissociation of oceanic methane hydrate as a cause of the carbon isotope excursion at the end of the Paleocene. *Paleoceanography* **10**, 965–971 (1995).
23. Katz, M. E., Pak, D. K., Dickens, G. R. & Miller, K. G. The source and fate of massive carbon input during the Latest Paleocene Thermal Maximum. *Science* **286**, 1531–1533 (1999).

24. Hesselbo, S. P. *et al.* Massive dissociation of gas hydrate during a Jurassic oceanic anoxic event. *Nature* **406**, 392–395 (2000).
25. Kvenvolden, K. A. in *Gas Hydrates: Relevance to World Margin Stability and Climate Change* (eds Henriot, J. P. & Mienert, J.) 9–30 (Spec. Publ. 137, Geological Society of London, 1998).
26. Hoffman, P. F., Halverson, G. P. & Grotzinger, J. P. Are Proterozoic cap carbonates and isotopic excursions a record of gas hydrate destabilization following Earth's coldest intervals? *Comment. Geology* **30**, 286–287 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank S. H. Zhang, H. C. Wu, G. B. Li and D. M. Liu for field assistance, and D. Mrofk and C. Valkenburg for laboratory assistance in stable isotopic analysis. We also thank M. A. Arthur and J. P. Grotzinger for comments and suggestions. This work was supported by the NSF.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.J. (ganqing@mail.ucr.edu).

A change in the freshwater balance of the Atlantic Ocean over the past four decades

Ruth Curry¹, Bob Dickson² & Igor Yashayaev³

¹Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA

²Centre for Environment, Fisheries, and Aquaculture Science, Lowestoft, NR33 0HT, UK

³Bedford Institute of Oceanography, Dartmouth, Nova Scotia, B2Y 4A2, Canada

The oceans are a global reservoir and redistribution agent for several important constituents of the Earth's climate system, among them heat, fresh water and carbon dioxide. Whereas these constituents are actively exchanged with the atmosphere, salt is a component that is approximately conserved in the ocean. The distribution of salinity in the ocean is widely measured, and can therefore be used to diagnose rates of surface freshwater fluxes¹, freshwater transport² and local ocean mixing³—important components of climate dynamics. Here we present a comparison of salinities on a long transect (50° S to 60° N) through the western basins of the Atlantic Ocean between the 1950s and the 1990s. We find systematic freshening at both poleward ends contrasted with large increases of salinity pervading the upper water column at low latitudes. Our results extend a growing body of evidence indicating that shifts in the oceanic distribution of fresh and saline waters are occurring worldwide in ways that suggest links to global warming and possible changes in the hydrologic cycle of the Earth.

The properties of Atlantic water masses have been changing—in some cases very much—over the five decades for which reliable and systematic records of ocean measurements are available. Careful analyses of observations have together revealed much about the magnitude, timing, and spatial scales of variability, and established a framework for interpreting the dynamics and kinematics underlying those changes^{4–10}. Building upon these studies and the lengthening timeline of observations, we have evaluated the time evolution of Atlantic salinities and find evidence for long-term and large-scale changes that appear to be organized, at least in part, around the structure of the hydrologic forcing as reflected in the evaporation minus precipitation (*E–P*) distribution.

We first present the large-scale differences in salinity that arose between the late 1950s and the 1990s along a particular transect

through the deep basins of the western Atlantic from 50° S to 60° N (black line in Fig. 1c). This line was chosen with some care. It spans the maxima and minima of *E–P* in both hemispheres^{11,12} (Fig. 1d). As shown, the *E–P* maxima underlying the trade-wind belts north and south of the Equator are separated by a belt of net precipitation associated with the intertropical convergence zone; a third *E–P* maximum follows the axis of the Gulf Stream. The main features of the surface salinity distribution result in large part from that *E–P* distribution, so our transect crosses the regions of maximum salinity in the subtropics of both hemispheres as well as the surface salinity minima along the Equator and at both poleward ends of the line (Fig. 1c).

Figures 1b and 2b show the differences in salinity observed along this transect for the time period 1985–99 compared to 1955–69 as functions of density and depth, respectively. The corresponding average distributions of salinity along this transect for the time period 1985–99 are also shown; all are annotated to indicate the locations of water mass types to which we shall refer. Along the section as a whole (Figs 1b and 2b), we find evidence of freshening over much of the water column at both poleward extremities of the section. In the intervening tropics and subtropics, we find that the upper ocean became more saline. These changes break down as follows.

North of 40° N the Atlantic water column became fresher by approximately –0.03 p.s.u. on average, reflecting the sustained freshening of LSW (see Fig. 1 legend for names of all water masses), but also of the products of Faroe–Shetland and Denmark Strait overflow from the Nordic seas which underlie it (that is, NEADW and DSOW)¹⁰.

Towards the southern limit of the transect, the water masses that outcrop in the regions where precipitation dominates—nominally south of 25° S in the western South Atlantic (see Fig. 1d)—have also freshened. Over the 40-yr record, the ventilated thermocline waters, in the neutral density range $\gamma_n = 25.5–27.0 \text{ kg m}^{-3}$, became less saline by more than –0.2 p.s.u.. The underlying AAIW and UCDW also show evidence of freshening, but at a comparatively reduced level of about –0.02 p.s.u.. A lack of deep-ocean measurements southward of 32° S in the earlier time frame precludes assessing changes below 3,000 m there.

In the waters of the tropics and subtropics, salinity increases observed over this period are greatest in the upper 500 m. Salinities increased by +0.1 to +0.4 p.s.u. over four decades in all Atlantic waters exposed to the atmosphere in the high-evaporation regions between 25° S and 35° N, corresponding to the density range $\gamma_n = 23.0–25.0 \text{ kg m}^{-3}$. Immediately underlying this layer, the subsurface thermocline waters in the density range $\gamma_n = 25.5–27.0 \text{ kg m}^{-3}$ became similarly more saline in the Northern Hemisphere. These are the SMW, whose properties are set at the sea surface in the eastern basin between 20° N and 30° N by hot, dry easterly winds from the African continent. From this source, SMW circulate westward in accordance with ventilated thermocline theory into the Caribbean and western North Atlantic¹³ where they are no longer in direct contact with the atmosphere, but are easily identified by a salinity maximum at depths of 100–300 m. In the South Atlantic, by contrast, maximum salinities are found at the sea surface on the western side of the basin (see Fig. 1c).

Although rising salinities are largely a feature of the low-latitude upper ocean, we find also some increase at depth between 25° S and 40° N. Centred on the neutral density level $\gamma_n = 27.8 \text{ kg m}^{-3}$, these elevated salinities correspond to the mixture of Atlantic and Mediterranean water masses that occupies depths 1,200–1,500 m, immediately overlying the UNADW. A 40-yr trend towards increased salinities has been documented in the deep waters of the Mediterranean Sea¹⁴ and the increase observed here reflects the trans-ocean spreading of that influence to the western boundary and southwards in the mid-depth Atlantic circulation. On our transect, salinity increases near $\gamma_n = 27.8 \text{ kg m}^{-3}$ exceed

+0.05 p.s.u. between 30°–40° N, the heart of the MOW influence at this longitude, while a lesser increase in salinity (+0.02 to +0.04 p.s.u.) can be traced at these densities southward to about 25° S.

Averaged vertically over the total water column, the observed changes in salinity—by approximately –0.03 p.s.u. north of latitude 40° N and +0.02 p.s.u. between 40° N and 25° S—are of remarkable amplitude. Taken together with the sign and pan-ocean structure of the changes, these results indicate that fresh water has been lost from the low latitudes and added at high latitudes, at a pace exceeding the ocean circulation's ability to compensate.

Although multiple factors have been implicated in these long-term changes, the available measurement record has not been sufficient to quantify their relative contributions to the observed trends, and that partitioning remains a high priority in ongoing research. The freshening of the entire system of overflow and entrainment that ventilates the deep North Atlantic has already been shown to have taken place at a remarkable, if not quite steady, rate of –0.010 to –0.015 p.s.u. per decade over the past four

decades¹⁰. That freshening has been attributed to some combination of enhanced wind-driven exports of ice or fresh water from the Arctic, increased net precipitation rates, and elevated volumes of continental runoff from melting ice^{15–18}—some of which can be associated with recent amplification of the North Atlantic Oscillation (NAO)¹⁹. In the low latitude Atlantic, the factors building positive salinity anomalies must also involve some combination of dynamic and thermo-dynamic processes: altered circulation, precipitation patterns and intensified trade winds²⁰—themselves associated with the NAO—but also enhanced evaporation rates due to warming of the surface ocean²¹.

We next examine the time history of zonally integrated salinity anomalies at 24° N—the latitude of the salinity maximum—for four density layers whose winter outcrop regions are shown in Fig. 3a. The time series in Fig. 3b documents a long-term trend of increasing salinity that bridges the time periods for which differences were shown in Figs 1 and 2. The positive anomaly is greatest (~+0.3 p.s.u.) in the 1990s and its largest contribution derives from the density layers that are most exposed to the winter

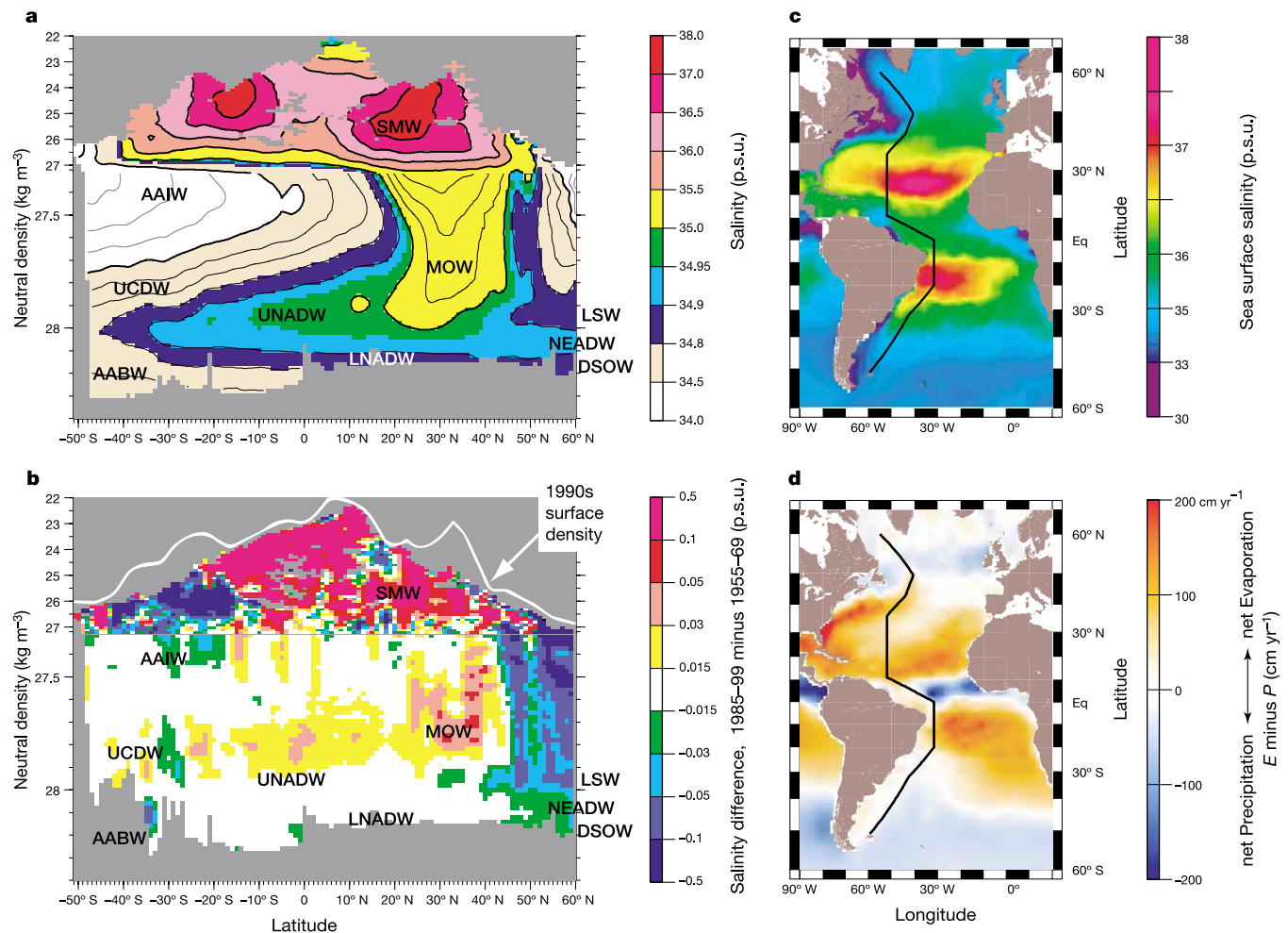


Figure 1 Western Atlantic meridional property sections and surface climatologies. **a**, Vertical section of salinity versus neutral density averaged for the years 1985–99. Grey shading means that no water at that density was measured. Annotations locate various water masses in all panels as follows. AAIW, Antarctic Intermediate Waters; UCDW, Upper Circumpolar Deep Waters; AABW, Antarctic Bottom Waters; UNADW, Upper North Atlantic Deep Waters; LNADW, Lower North Atlantic Deep Waters; MOW, Mediterranean Outflow Waters; SMW, Salinity Maximum Waters; LSW, Labrador Sea Water; NEADW, Northeast Atlantic Deep Water; DSOW, Denmark Strait Overflow Water. **b**, Salinity difference (1985–99 minus 1955–69) versus neutral density. Grey shading means that no water at that

density was measured in one or both of the two time periods. No deep measurements were available south of 32° S in the earlier time frame. Sea surface densities had become considerably lighter in 1985–1999 compared to 1955–69, as depicted by the white line. No salinity difference can be computed for those surface densities, but changes at the sea surface are shown versus depth in Fig. 2. Upper-ocean density decreased in the Northern Hemisphere despite salinity increases >0.4 p.s.u., indicating that warming of the upper ocean was dominating the density changes. **c**, Climatological sea surface salinity. The black line delineates the location of the meridional section. **d**, Annual average $E-P$ (courtesy of R. Schmitt, WHOI). p.s.u., practical salinity units.

atmosphere in the high $E-P$ region along that section, that is $\gamma_n = 25.5\text{--}26.0$ and $26.0\text{--}26.5 \text{ kg m}^{-3}$. The deeper density layers, $26.5\text{--}27.0$ and $27.0\text{--}27.5 \text{ kg m}^{-3}$, which outcrop northward of 30°N and 40°N respectively—away from the $E-P$ maximum—show much smaller salinity changes ($<+0.1$ p.s.u.) over the total span of the record.

Because salt is neither gained nor lost through the atmosphere, a salinity increase in a layer implies either that additional salt (or less freshwater) was mixed in from surrounding waters, or that extra fresh water was removed. Because SMW are the most saline waters in the North Atlantic, no source for additional salt exists. Furthermore, the underlying layers exhibit no compensating freshening, so that local ocean mixing can be ruled out as the cause of the observed salinity increases. We have therefore evaluated the annual freshwater loss that could be attributed to local $E-P$ changes for each of two

layers, both of which are bounded above by the sea surface. The first layer considered is bounded below by the density surface $\gamma_n = 26.50 \text{ kg m}^{-3}$, which includes all waters that are ventilated in the climatological $E-P$ maximum south of 30°N . The second layer is bounded by a somewhat deeper surface, $\gamma_n = 27.0 \text{ kg m}^{-3}$, which outcrops as far north as 40°N .

The time series of estimated $E-P$ anomaly, shown in Fig. 3c, describes a basin-wide trend of enhanced freshwater loss with time in the maximum net evaporation region along 24°N . The slope of the line fitted to the time series defines the rate of estimated $E-P$ change in cm yr^{-1} . Viewed as a long-term trend, the data indicate a 40-yr $E-P$ increase of approximately 5 cm yr^{-1} . Alternatively, the rate of change could be characterized as an initial rise to a 20-yr plateau of $\sim 50 \text{ cm}$, followed by a $\sim 10 \text{ cm yr}^{-1}$ increase in the 1990s. In either scenario, an additional 150–200 cm of fresh water has been

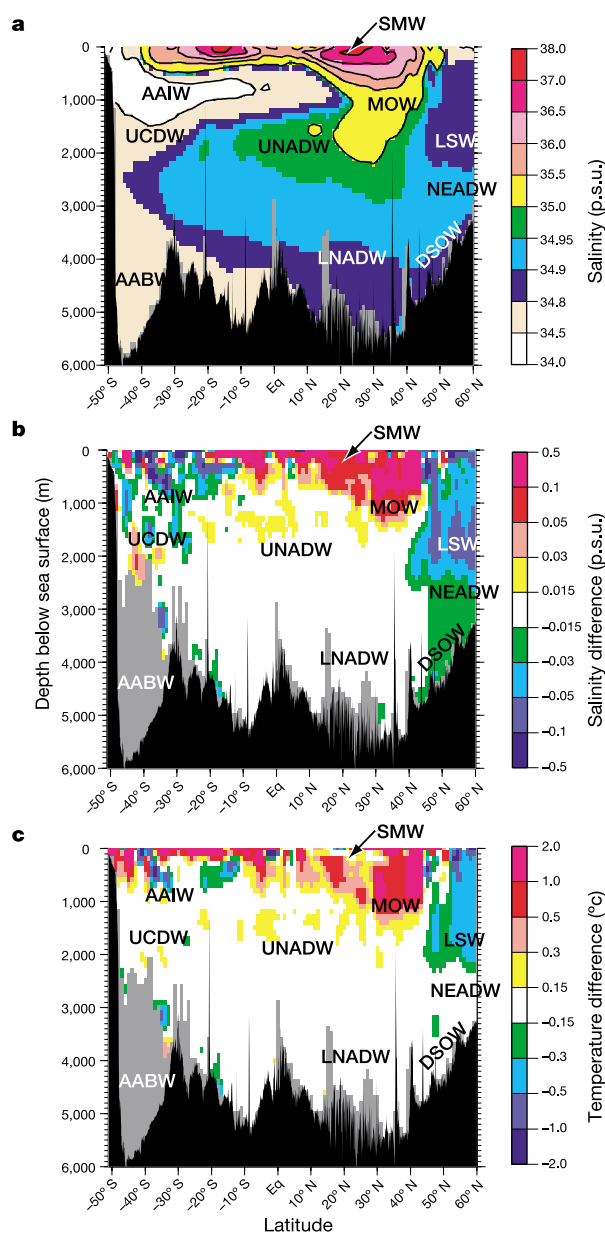


Figure 2 Western basin meridional property sections versus depth. **a**, Vertical section of salinity versus depth averaged for the years 1985–99. Water mass locations are marked as in Fig. 1. **b**, Salinity difference (1985–99 minus 1955–69) versus depth. **c**, Temperature difference versus depth for the same time periods.

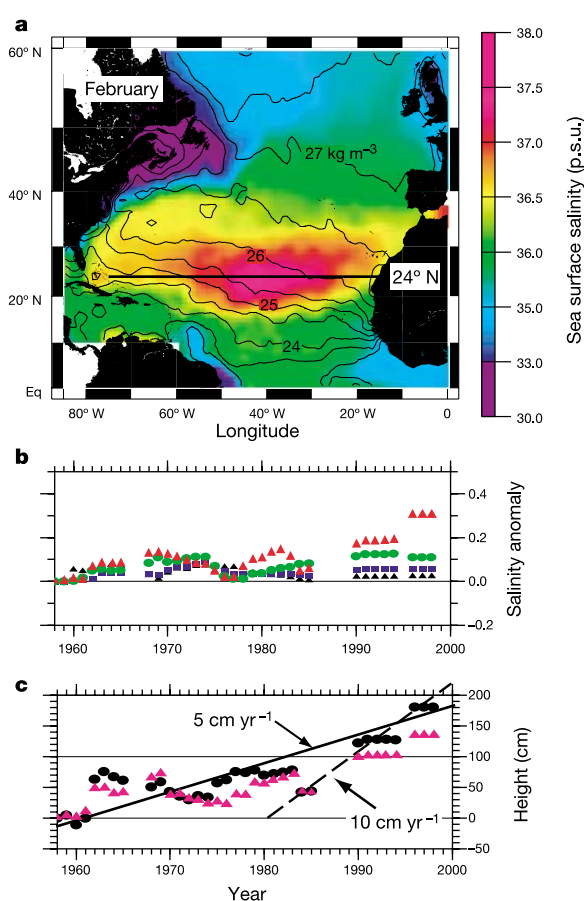


Figure 3 Time series of upper ocean salinity anomalies and estimated $E-P$ anomalies. **a**, Map of February mean salinity (colour scale) and neutral density (contour interval is 0.5 kg m^{-3}) at the sea surface. The 24°N section from which time series were constructed is shown. **b**, Time series of salinity anomaly (relative to the first year in the time series) integrated across 24°N for four neutral density layers, $25.5\text{--}26.0 \text{ kg m}^{-3}$ (red triangles), $26.0\text{--}26.5 \text{ kg m}^{-3}$ (green circles), $26.5\text{--}27.0 \text{ kg m}^{-3}$ (blue squares), and $27.0\text{--}27.5 \text{ kg m}^{-3}$ (black triangles). Approximate winter outcrop regions for each layer are shown in **a**. **c**, Estimated $E-P$ anomaly for two layers: sea surface to $\gamma_n = 26.5 \text{ kg m}^{-3}$ (magenta triangles), and sea surface to $\gamma_n = 27.0 \text{ kg m}^{-3}$ (black circles). Each symbol indicates the amount of extra fresh water that would have to be removed per unit area along the section to account for the observed change in salinity of that layer in each year. The slope of the line fitted to the black data points provides an estimate for the rate of change of $E-P$ in cm yr^{-1} . The solid line estimates the $E-P$ anomaly for the entire 42-year record span; the dashed line represents the rate of change estimated for the last 15 years. The y-axis is the height of a column of fresh water removed by evaporation to account for the salinity change.

lost per unit area in the 1990s relative to the late 1950s. Because climatological $E-P$ values are of the order of 100 cm yr^{-1} across this section, this represents a 5–10% increase in net evaporation.

The inferred 1990s rise in evaporation rate coincided with a protracted high state of the NAO. The associated increase in trade winds will have enhanced both evaporation and Ekman pumping of the SMW into the western basin's ventilated thermocline²². But a growing body of evidence also raises the possibility of a link to global warming and the planetary hydrologic cycle. First, tropical and subtropical upper ocean temperatures have risen in the Atlantic over the past few decades^{23–25}. Temperature differences evaluated along our transect (Fig. 2c) indicate an extensive warming of the upper ocean (about 1°C , but with the caveat that the data are not sufficient to resolve the annual heating cycle). Moreover, there is an unambiguous connection between the water masses that became more saline (Fig. 2b) and those that warmed, a feature which is confirmed by similar analyses applied to a transect through the Atlantic eastern basins. To this we add that the Clausius–Clapeyron equation²¹ predicts a 5–10% increase of water vapour pressure for a temperature rise of $0.5\text{--}1.0^\circ\text{C}$ in the range $20\text{--}25^\circ\text{C}$ —similar to the order of changes in net evaporation rates we have estimated along 24°N .

In addition, parallel changes in ocean salinity and temperature distributions are occurring in other oceans. In the Mediterranean, as previously mentioned, the water masses formed by net evaporation exhibit a 40-yr trend of increasing temperature and salinity¹⁴. In the Pacific, a symmetric freshening of intermediate waters ventilated at high latitudes of the Northern and Southern hemispheres contrasts with zonally averaged salinity and temperature increases in the upper 200 m at several lines spanning latitudes 24°N to 32°S (ref. 26). In a subsequent analysis, the possibility that such coherent behaviour might reflect some shorter-term natural oscillation in the Pacific climate, such as the Pacific Decadal Oscillation (PDO), was discounted on the basis that recent freshening of AAIW had also been observed in the Indian Ocean where the PDO signal is slight^{27,28}. An analogous argument could be extended to the hemispherically symmetric Atlantic salinity changes, to downplay the NAO as the sole dynamic at work.

Although it is a fundamental component of the planetary energy budget, the hydrologic cycle remains one of the least-understood elements of the climate system and freshwater budgets one of the largest causes for differences among climate models²⁹. Given the great uncertainties in measuring evaporation and precipitation over the oceans, conclusive evidence for changes in the global water cycle will depend on present and future efforts to directly measure salinity changes and freshwater transports by ocean currents. □

Methods

Climatologies

Ocean climatologies were constructed using the HydroBase2 package³⁰ and its database of hydrographic profiles, which have been rigorously screened for quality. Three-dimensional fields of salinity, temperature, and neutral density at 1° grid resolution were generated for various time frames (5-, 10- and 15-yr composites) using isopycnal gridding and interpolation methods. Changes in hydrographic properties between time periods were evaluated as a function of both density and depth by subtracting identical grid points along a vertical section extracted from each climatology. Uncertainties for the data varied with time and are estimated as follows. For 1955–1969, $T = \pm 0.01^\circ\text{C}$ and $S = \pm 0.01 \text{ p.s.u.}$. For 1970–1989, $T = \pm 0.005^\circ\text{C}$ and $S = \pm 0.005 \text{ p.s.u.}$. For 1990–1999, $T = \pm 0.001^\circ\text{C}$ and $S = \pm 0.002 \text{ p.s.u.}$. Lower limits for a significant difference in salinity was set at $\pm 0.015 \text{ p.s.u.}$ and for temperature at $\pm 0.15^\circ\text{C}$. All of the upper-ocean signals reported here are an order of magnitude greater (0.1–0.4) than the worst-case error; and the deep-ocean anomalies exceed twice the measurement uncertainties.

Layer-averaged salinity and $E-P$ anomalies

Five-year running time frames spanning the instrumental record (1955–2000) were used to estimate yearly climatologies (a 1-2-5-2-1 weighting scheme was implemented for each time frame). Time series of annual layer-averaged salinity values, S_i , for the 24°N sections were evaluated by:

$$S_i = \frac{\int \rho_s dz dx}{\int \rho dz dx} \quad (1)$$

where s is salinity (in p.s.u.), ρ is *in situ* density (in kg m^{-3}), dz is layer thickness (in m), and dx is the distance (in m) between successive profiles along the section. Temporal changes were defined by the difference between S_i and S_1 , the first year in each time series (1958). Years for which the section contained spatial gaps exceeding 100 km after interpolation were omitted from the time series.

The freshwater loss that could be attributed to local $E-P$ changes was evaluated by invoking an approximate salt conservation statement:

$$\rho_1 S_1 H = \rho_2 S_2 (H + \Delta h) \quad (2)$$

where H is the average thickness of a layer (in cm) and Δh is the height (in cm) of fresh water removed through the air–sea boundary per unit area for each section.

Received 11 September; accepted 14 November 2003; doi:10.1038/nature02206.

1. Wüst, G. *Länderkundliche Forschung* 347–359 (Festschrift Norbert Krebs, Stuttgart, 1936).
2. Wüffels, S., Schmitt, R., Bryden, H. & Stigebrandt, A. Transport of freshwater by the oceans. *J. Phys. Oceanogr.* **22**, 155–162 (1992).
3. Lukas, R. & Lindstrom, E. The mixed layer of the Western Equatorial Pacific Ocean. *J. Geophys. Res.* **96** (suppl.), 3343–3357 (1991).
4. Levitus, S. Interpentadal variability of salinity in the upper 150m of the North Atlantic Ocean, 1970–1974 versus 1955–1959. *J. Geophys. Res.* **94**, 9679–9685 (1989).
5. Deser, C. & Blackmon, M. Surface climate variations over the North Atlantic Ocean during winter: 1900–1989. *J. Clim.* **6**, 1743–1753 (1993).
6. Bryden, H. et al. Decadal changes in water mass characteristics at 24°N in the subtropical North Atlantic Ocean. *J. Clim.* **9**, 3162–3186 (1996).
7. Joyce, T., Pickart, R. & Millard, R. Long-term hydrographic changes at 52° and 66°W in the North Atlantic subtropical gyre & Caribbean. *Deep Sea Res.* **46**, 245–278 (1999).
8. Lazier, J. et al. *Natural Climate Variability on Decade-to-Century Time Scales* (ed. Martinson, D. G.) 295–304 (National Academy Press, Washington DC, 1995).
9. Blindheim, J. et al. Upper layer cooling and freshening in the Norwegian Sea in relation to atmospheric forcing. *Deep-Sea Res.* **47**, 655–680 (2000).
10. Dickson, R. R., Yashayaev, I., Meincke, J., Turrell, W., Dye, S. & Høffert, J. Rapid freshening of the Deep North Atlantic over the past four decades. *Nature* **416**, 832–837 (2002).
11. Schmitt, R., Bogden, P. & Dorman, C. Evaporation minus precipitation and density fluxes for the North Atlantic. *J. Phys. Oceanogr.* **10**, 1210–1221 (1989).
12. Schmitt, R. W. The ocean component of the water cycle. *Rev. Geophys.* **33**, 1395–1409 (1995).
13. Blanke, B., Arhan, M., Lazar, A. & Prevost, G. A Lagrangian numerical investigation of the origins and fates of the Salinity Maximum Water in the Atlantic. *J. Geophys. Res.* **107**, 1–15 (2003).
14. Roether, W. et al. Recent Changes in the Eastern Mediterranean Deep Waters. *Science* **271**, 333–335 (1996).
15. Dickson, R. R. et al. The Arctic Ocean response to the North Atlantic Oscillation. *J. Clim.* **13**, 2671–2696 (2000).
16. Vinje, T. Fram Strait ice fluxes and atmospheric circulation, 1950–2000. *J. Clim.* **14**, 3508–3517 (2001).
17. Peterson, B. J. et al. Increasing river discharge to the Arctic Ocean. *Science* **298**, 2171–2173 (2002).
18. Abdalati, W. & Steffen, K. Snowmelt on the Greenland ice sheet as derived from passive microwave satellite data. *J. Clim.* **10**, 165–175 (1997).
19. Marshall, J. et al. North Atlantic Climate Variability: phenomena, impacts and mechanisms. *Int. J. Clim.* **21**, 1863–1898 (2001).
20. Moulin, C., Lambert, C. E., Dulac, F. & Dayan, U. Control of atmospheric export of dust from North Africa by the North Atlantic Oscillation. *Nature* **387**, 691–694 (1997).
21. Peixoto, J. P. & Oort, A. H. *Physics of Climate* 53 (American Institute of Physics, New York, 1992).
22. Mignot, J. & Frankignoul, C. On the interannual variability of surface salinity in the Atlantic. *Clim. Dyn.* **20**, 555–565 (2003).
23. Levitus, S., Antonov, J., Boyer, T. & Stephens, C. Warming of the World Ocean. *Science* **287**, 2225–2229 (2000).
24. Joyce, T. & Robbins, P. The long-term hydrographic record at Bermuda. *J. Clim.* **9**, 3121–3131 (1996).
25. Parilla, G., Lavin, A., Bryden, H., Garcia, M. & Millard, R. Rising temperatures in the subtropical North Atlantic Ocean over the past 35 years. *Nature* **369**, 48–51 (1994).
26. Wong, A. P. S., Bindoff, N. L. & Church, J. L. Large scale freshening of intermediate waters in the Pacific and Indian Oceans. *Nature* **400**, 440–443 (1999).
27. Wong, A. P. S., Bindoff, N. L. & Church, J. C. Freshwater and heat changes in the North and South Pacific Oceans between the 1960s and 1985–94. *J. Clim.* **14**, 1613–1633 (2001).
28. Bindoff, N. L. & McDougall, T. J. Decadal changes along an Indian Ocean Section at 32°S and their interpretation. *J. Phys. Oceanogr.* **30**, 1207–1222 (2000).
29. Intergovernmental Panel on Climate Change (IPCC). *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the IPCC* (eds Houghton, J. T. et al.) (Cambridge Univ. Press, Cambridge, UK, 2001).
30. Curry, R. G. *HydroBase2: A Database of Hydrographic Profiles and Tools for Climatological Analysis* (<http://www.whoi.edu/science/PO/hydrobase>) (WHOI, 2002).

Acknowledgements We are grateful to J. Toole and R. Schmitt for their encouragement and discussions. This work was supported by an NSF grant and the WHOI Ocean and Climate Change Institute. It has also formed part of the data synthesis phase of the WOCE Program, of the ASOF Project of the US NSF and the Framework V Programme of the European Community, and of the NOAA Consortium on the Ocean's Role in Climate of the Scripps Institution of Oceanography and the Lamont–Doherty Earth Observatory.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.C. (rccurry@whoi.edu).

Palaeolithic ivory sculptures from southwestern Germany and the origins of figurative art

Nicholas J. Conard

Abteilung für Ältere Urgeschichte und Quartärökologie, Institut für Ur- und Frühgeschichte und Archäologie des Mittelalters, Universität Tübingen, Schloss Hohentübingen, 72070 Tübingen, Germany

Archaeologists have always viewed the origin of figurative art as a crucial threshold in human evolution^{1,2}. Here I report the discovery of three figurines carved from mammoth ivory at Hohle Fels Cave in the Swabian Jura of southwestern Germany, which provides new evidence for the appearance of figurative art more than 30,000 years ago. The finds include the oldest known representation of a bird, a therianthrope sculpture and an animal that most closely resembles a horse. The Aurignacian sculptures of the Swabian Jura belong to one of the oldest traditions of figurative art known worldwide and point to the Upper Danube as an important centre of cultural innovation during the early Upper Palaeolithic period^{3,4}.

Hohle Fels is located in the Ach Valley near the town of Schellklingen, 20 km southwest of Ulm. After many years of excavation in the rich Magdalenian and Gravettian deposits, in 1999 the excavation team reached the Aurignacian strata⁴. There, along with a wealth of lithic and organic artefacts, three figurines of mammoth ivory have been recovered (Figs 1 and 2). These figurines complement the assemblages of ivory sculptures found in Aurignacian cave deposits at the Swabian sites of Vogelherd, Hohlenstein-Stadel and Geißenklösterle (ref. 3 and Table 1).

Figurine 1 depicts the head of an animal that is most probably a horse, although it could perhaps represent a bear or another animal. The main piece was found in 1999 in the transition between archaeological horizons (AH) IId and Ile, and it fits to a piece of the animal's cheek from the underlying AH IIIa. The sides of the face and underside of the jaw show fine, regular cross hatching and fine parallel lines. The mandibular morphology is accentuated by a line along the base of the jaw, and the mouth, nostrils and eyes of the animal are depicted with deeply incised lines.

The body of the Figurine 2 was recovered in 2001 from AH IV near the bottom of the Aurignacian sequence. In 2002 the head and neck of the figurine were recovered from AH IV and confirm that the sculpture depicts a water bird with a morphology similar to that of a diver, cormorant or duck. This figurine has dimensions of 47 × 13 × 9 mm. The extended neck of the bird is strongly suggestive of a waterfowl in flight or diving. The wings are depicted close to the body. As compared with many finely carved figurines of the Swabian Aurignacian, the front of the bird has been left in a seemingly unfinished state. The other parts of the bird are presented in greater detail. Both eyes are easily recognizable, and the beak has a conical, pointy form that one would not expect on many of the common ducks. The legs of the bird are short with no indications of feet. The tail of the figurine extends below the legs and is depicted as a finely carved flat splint. The back of the bird shows a series of distinct lines that apparently represents feathers.

Figurine 3 stands only 25.5 mm high and was recovered in 2002 from AH IV. The figurine is a longitudinal fragment of the head, torso, arm, shoulder and buttocks of an upright therianthrope representation, showing marked similarities to the *Löwenmensch* from Hohlenstein-Stadel. The shoulder is angular and the posture rigid. A delicately carved ear is visible high on the head. The arm is short and pointy with an incised vertical line. The main features

depict a mixture of felid and human traits. The preservation of the piece allows no conclusions to be drawn about the sexual attribution of the figurine.

The horse head is slightly older than 30,000 years (30 kyr), the figurines from AH IV date to 31–33 kyr, and the deepest Aurignacian deposit, AH Va, which immediately underlies AH IV, has yielded dates between 33 and 36 kyr (Table 2). Radiocarbon dates for the period before 30 kyr are likely to be underestimates of calendar ages^{4–7}. Because there is no reliable, high-resolution radiocarbon calibration for this period, the precise ages of the Swabian ivory artworks are not known.

Archaeological horizon IV is the richest of the Aurignacian deposits at Hohle Fels and, although only 9 m² have been excavated, the deposit has provided a rich assemblage of lithic and organic artefacts, including diverse forms of finely carved ivory ornaments and much ivory working debris. Whereas the setting of the discovery of the original *Löwenmensch* deep in the cave of Hohlenstein-Stadel has been argued to be a cache or cult site^{3,8}, Vogelherd, Geißenklösterle and Hohle Fels, on the basis of the size and content

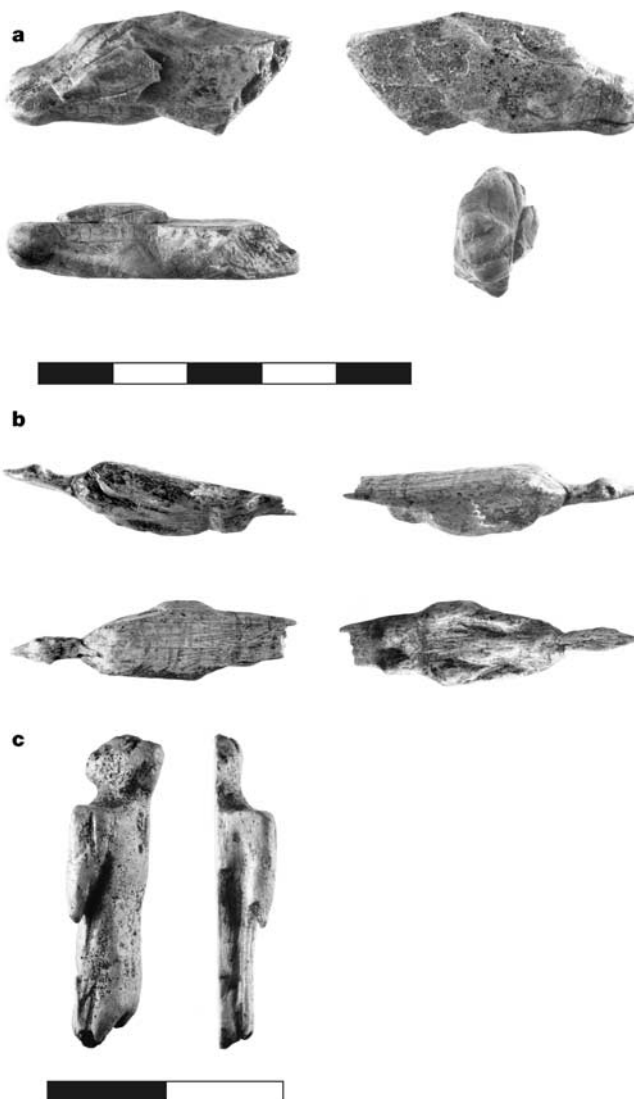


Figure 1 Views of the three ivory figurines from Hohle Fels, depicting the head of a horse (a), a water bird (b) and a therianthrope with the characteristics of a felid and human (c). Scale bars, 5 cm (a, b); 2 cm (c). Photos H. Jensen, copyright University of Tübingen.

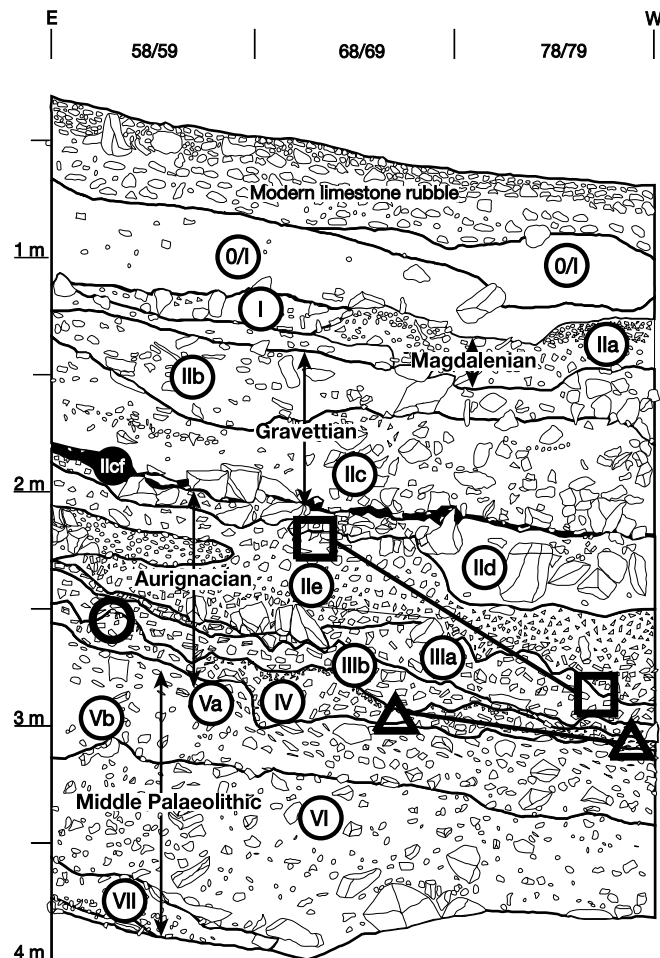


Figure 2 Stratigraphic provenience of the three Hohle Fels figurines. The horse head and bird comprise two refitted pieces. Triangle, bird; circle, therianthrope; square, horse head.

of their assemblages, represent substantial habitation sites. Work on the fauna from the Ach Valley sites indicates that the caves of the region were used repeatedly in the winter and spring for relatively lengthy occupations⁹. Some of these horizons have produced hundreds of pieces of debris from ivory working.

Over the years there has been considerable debate and specu-

lation about the meaning of these figurines. Riek, for example, emphasized the importance of palaeoecology and hunting magic¹⁰, whereas Hahn, in his thesis "*Kraft und Aggression*", argued that the Aurignacian inhabitants of the region mainly depicted strong, fast and dangerous animals³. Lewis-Williams, Porr and others have stressed the importance of mixed representations of animals and humans as evidence for shamanism^{2,11}.

The new figurines place some constraints on the existing range of interpretations. Although the discovery of another horse and a *Löwenmensch* may be viewed as consistent with the *Kraft und Aggression* hypothesis, the waterfowl from Hohle Fels is not consistent with this interpretation. One increasingly gains the impression that the Swabian Aurignacian figurines, while emphasizing predators as well as large, strong and fast animals, also depict a broad range of animals that the occupants of the region presumably admired.

Particularly illuminating is the discovery of a second *Löwenmensch*. The occupants of Hohle Fels in the Ach Valley and Hohlenstein-Stadel in the Lone Valley must have been members of the same cultural group and shared beliefs and practices connected with therianthrope images of felids and humans. These figures span the complete range of sizes for the carved ivory depictions. Whereas the original find is 30 cm high and originally weighed about 750 g, the find from Hohle Fels originally stood less than 4 cm high and weighed only a few grams. The discovery of a second *Löwenmensch* lends support to the hypothesis that Aurignacian people practised a form of shamanism.

The ivory figurines from Swabia represent one of the earliest artistic traditions worldwide. In Africa, the earliest evidence for figurative representations is the mobile rock art from late MSA contexts at Apollo 11 in southwestern Namibia, which dates to between 25.5 and 27.5 kyr (ref. 12). These paintings depict animals and a possible therianthrope. At Fumane Cave in northern Italy, preliminary information indicates that paintings including representations in red pigments of a therianthrope and several animals have been found in Aurignacian contexts predating 30 kyr (ref. 13). In Asturias in northern Spain at the site of Candamo, controversial claims for wall paintings have been made for the period around 32 kyr (refs 14, 15). In southwestern France at Abri Cellier, La Ferrassie, Abri Blanchard and Abri Castanet, there is evidence for engravings and paintings including representations of vulvas and animals dating to ~30 kyr (ref. 1). In Stratzing in Lower Austria, a small human figurine of stone has been recovered from Aurignacian contexts dating between 30 and 32 kyr (ref. 16).

The most spectacular evidence for early figurative artwork out-

Table 1 Ivory figurines from the Aurignacian of the Swabian Jura

Site	Figurine	Archaeological horizons	Date ¹⁴ C (kyr)	Date TL (kyr)	Excavation, date
Hohle Fels	Horse	IId/Ile-IIIa	30–31		Conard, Uerpmann 1999
	Bird	IV	31–33		Conard, Uerpmann 2001 and 2002
	'Löwenmensch'	IV	31–33		Conard, Uerpmann 2002
Geißenklösterle	Mammoth	Ila	30–34	~37	Hahn 1974
	Bear	Ila	30–34	~37	Hahn 1974 and 1977
	Anthropomorphic relief	IIb	30–34	~37	Hahn 1979
	Bovid	IIb	30–34	~37	Hahn 1983
Hohlenstein-Stadel	'Löwenmensch'	20 m, spit 6	31–32		Wetzel, Völzing 1939
Vogelherd	Anthropomorph	IV	~30–32		Riek 1931
	Felid	IV	~30–32		Riek 1931
	Bovid	IV	~30–32		Riek 1931
	Mammoth	V	30–36		Riek 1931
	Mammoth	V	30–36		Riek 1931
	Horse	V	30–36		Riek 1931
	Felid	V	30–36		Riek 1931
	Felid	V	30–36		Riek 1931
	Unidentified quadruped	V	30–36		Riek 1931
	Felid	Backdirt	–		Weber 1952, Scheer before 1971

Several artworks that are not figurative representations in ivory are not included in the table.

Table 2 AMS radiocarbon dates from Gravettian and Aurignacian deposits at Hohle Fels*

Laboratory number	Archaeological horizons	Material	Modification	Date (years)	Cultural group
OxA-4599	IIc	Reindeer antler	Tool (decorated adze)	28,920 ± 400	Gravettian
OxA-5007	IIc	Reindeer antler	Tool (decorated adze)	29,550 ± 650	Gravettian
KIA 3503	IIcf	Bone		27,030 ± 250 – 240	Gravettian
KIA 17742	IIcf	Horse tibia	Impact + cut marks	27,690 ± 140	Gravettian
KIA 17743	IIcf	Cave bear vertebra	With flint point	27,830 ± 150/–140	Gravettian
KIA 17744	IIcf	Mammoth/rhino rib	Point	27,780 ± 150	Gravettian
KIA 17741	IIcf	Reindeer antler		27,970 ± 140	Gravettian
KIA 8964	IIId (base)	Mammoth/rhino rib		29,560 ± 240/–230	Aurignacian
KIA 8965	IIId (base)	Reindeer antler		30,010 ± 220	Aurignacian
KIA 16040	IIe	Horse pelvis	Impact + cut marks	30,640 ± 190	Aurignacian
KIA 16038	IIId	Reindeer femur	Impact + cut marks	29,840 ± 210	Aurignacian
KIA 18877	IIId	Pinus charcoal		30,170 ± 250/–240	Aurignacian
OxA-4601	IIId	Bone		30,550 ± 550	Aurignacian
KIA 18876	IIId	Pinus charcoal		31,010 ± 600/–560	Aurignacian
KIA 16039	IIId	Small ungulate femur	Impact	31,140 ± 250/–240	Aurignacian
KIA 18878	IIId	Pinus charcoal		29,780 ± 330/–310	Aurignacian
OxA-4600	IV	Reindeer metapodial		31,100 ± 600	Aurignacian
KIA 18879	IV	Unidentified charcoal		31,160 ± 1,530/–1280	Aurignacian
KIA 16036	IV	Horse femur	Tool (retoucher)	33,090 ± 260/–250	Aurignacian
KIA 16035	Va	Horse bone	Impact	33,290 ± 270	Aurignacian
KIA 18880	Va	Pinus charcoal		34,190 ± 340/–330	Aurignacian
KIA 16034	Va	Small ungulate humerus	Impact + cut marks	35,710 ± 360/–340	Aurignacian

*AMS, accelerator mass spectrometry.

side Swabia comes from Grotte Chauvet in the Ardèche region of southern France, where rock paintings have been published dating as far back as 32 kyr (refs 15,17–19). Debate continues on the antiquity of the paintings from Grotte Chauvet, and similarities between the subject matter depicted at Chauvet and in the Swabian Aurignacian have been taken as evidence for their antiquity. Fully symbolic communication and cultural modernity may well have existed earlier in other regions, but nowhere is the concrete evidence for cultural modernity in the form of numerous figurative depictions combined with the regular manufacture and use of ornaments and the presence of musical instruments demonstrated earlier than in Swabia⁴.

Because most Aurignacian deposits have produced no diagnostic human remains, there is no proof as to which hominid created these assemblages²⁰. If, however, one assumes that modern humans produced these assemblages, the Aurignacian of Swabia suggests an early migration of modern humans along the Danube Corridor into the interior of the continent⁴. The Aurignacian appears suddenly in Swabia, and nowhere in Swabia is an interstratification of Middle Palaeolithic and Aurignacian assemblages documented. At several key sites an occupational hiatus is recorded in the strata below the Aurignacian. The “population vacuum” hypothesis^{4,21} postulates that modern humans migrated via the Danube Corridor into Swabia immediately after a cold arid period, most probably the terrestrial equivalent of the Heinrich 4 event dating to ~38–40 kyr (refs 22, 23). This harsh climatic phase ended suddenly with the start of Dansgaard–Oeschger cycle IS 8, which dates to roughly 38 kyr. This could explain why modern humans apparently arrived in an empty or largely depopulated region. Early Upper Palaeolithic migrations of modern humans also occurred along the Mediterranean coast and other routes. Innovations spread rapidly and regional variation ensued^{24,25}. The Swabian Jura preserves outstanding evidence for the early Aurignacian, and is one of several key areas of cultural florescence at the beginning of the Upper Palaeolithic. □

Received 18 September; accepted 7 November 2003; doi:10.1038/nature02186.

1. Lewis-Williams, D. *The Mind in the Cave* (Thames and Hudson, London, 2002).
2. Leroi-Gourhan, A. *Préhistoire de l'Art Occidental* (Citadelles & Mazenod, Paris, 1995).
3. Hahn, J. *Kraft und Aggression. Die Botschaft der Eisezeitkunst im Aurignacien Süddeutschlands?* (Archaeologica Venatoria, Tübingen, 1986).
4. Conard, N. J. & Bolus, M. Radiocarbon dating the appearance of modern humans and the timing of cultural innovations in Europe: new results and new challenges. *J. Hum. Evol.* **44**, 331–371 (2003).
5. Voelker, A. H. L., Grootes, P. M., Nadeau, M.-J. & Sarntinhe, M. Radiocarbon levels in the Iceland

Sea from 25–53 kyr and their link to the earth's magnetic intensity. *Radiocarbon* **42**, 437–452 (2000).
6. Beck, J. W. *et al.* Extremely large variations of atmospheric ¹⁴C concentration during the Last Glacial period. *Science* **292**, 2453–2458 (2001).
7. Richter, D., Waiblinger, J., Rink, W. J. & Wagner, G. A. Thermoluminescence, electron spin resonance and ¹⁴C-dating of the late Middle Paleolithic and the early Upper Paleolithic site of Geißenklösterle in southern Germany. *J. Archaeol. Sci.* **27**, 71–89 (2000).
8. Schmid, E. Die altsteinzeitliche Elfenbeinstatue aus der Höhle Stadel im Hohlenstein bei Asselfingen, Alb-Donau-Kreis. *Fundb. Baden Württemb* **14**, 33–118 (1989).
9. Münzel, S. C. Cave bear hunting in the Swabian Alb (Germany), 30,000 years ago. *Abhandl. Karst- und Höhlenkunde* **34**, 36–39 (2002).
10. Riek, G. *Die Eizeitjägerstation am Vogelherd im Lonetal I.* (Heine, Tübingen, 1934).
11. Por, M. in *Eiszeitkunst im Süddeutsch-Schweizerischen Jura* (eds Müller-Beck, H., Conard, N. J. & Schürle, W.) 95–102 (Theiss, Stuttgart, 2001).
12. Vogelsang, R. *Middle-Stone-Age-Fundstellen in Südwest-Namibia* (Heinrich Barth Institut, Cologne, 1998).
13. Broglio, A. Paleolithic e Mesolithic. *Preistoria Veronese* **5**, 11–56 (2002).
14. Fortea, J. Los comienzos del arte paleolítico en Asturias: aportaciones desde una arqueología contextual no postestilística. *Zephyrus* **53**, 177–216 (2001).
15. Pettitt, P. & Bahn, P. Current problems in dating Palaeolithic cave art: Candamo and Chauvet. *Antiquity* **77**, 134–141 (2003).
16. Neugebauer-Maresch, C. Zum Neufund einer weiblichen Statuette bei den Rettungsgrabungen an der Aurignacien-Station Stratzing/Krems-Rehberg, Niederösterreich. *Germania* **67**, 551–559 (1989).
17. Clottes, J. (ed.) *La Grotte Chauvet: L'Art des Origines* (Seuil, Paris, 2001).
18. Valladas, H. & Clottes, J. Style, Chauvet and radiocarbon. *Antiquity* **77**, 142–145 (2003).
19. Züchner, C. La cueva Chauvet, datada arqueológicamente. *Edades Revista Historia* **6**, 167–185 (1999).
20. Churchill, S. E. & Smith, F. H. Makers of the Early Aurignacian of Europe. *Yearb. Phys. Anthropol.* **43**, 61–115 (2001).
21. Conard, N. J., Langguth, K. & Uerpmann, H.-P. Einmalige Funde aus dem Aurignacien und erste Belege für ein Mittelpaläolithikum im Hohle Fels bei Schelklingen, Alb-Donau-Kreis. *Archäol. Ausgr. Baden Württemb.* 21–27 (2003).
22. Dansgaard, W. *et al.* Evidence for general instability of past climate from 250-kyr ice-core record. *Nature* **364**, 218–220 (1993).
23. Stuiver, M. & Grootes, P. M. GISP2 oxygen isotope ratios. *Quatern. Res.* **53**, 277–284 (2000).
24. Hahn, J. *Aurignacien. Das ältere Jungpaläolithikum in Mittel- und Osteuropa* (Böhlau, Cologne, 1977).
25. Bon, F. *L'Aurignacien entre Mer et Océan: Réflexion sur l'Unité des Phases Anciennes de l'Aurignacien dans le Sud de la France Mémoire* 29 (Soc. Préhist. Franc, 2002).

Acknowledgements Many colleagues working on the Palaeolithic archaeology and palaeoecology of the Swabian Jura have contributed to this research, including but not limited to M. Bolus, H. Floss, P. Goldberg, H. Jensen, A. Kandel, P. Krönneke, K. Langguth, M. Malina, U. Müller, S. Münzel, L. Niven, D. Richter, F. Smith, H.-P. Uerpmann and K. Wehrberger. I thank M. Collard for comments. This research has been supported by the Deutsche Forschungsgemeinschaft, the University of Tübingen, the Landesdenkmalamt Baden-Württemberg, the Alb-Donau-Kreis, Heidelberger Cement, the Museumsgesellschaft Schelklingen and the Gesellschaft für Urgeschichte.

Competing interests statement The author declares that he has no competing financial interests.

Correspondence and requests for materials should be addressed to the author (nicholas.conard@uni-tuebingen.de).

A larval Devonian lungfish

Keith S. Thomson, Mark Sutton & Bethia Thomas

Oxford University Museum of Natural History, Parks Rd, Oxford OX1 3PW, UK

Perhaps the most enduring of puzzles in palaeontology has been the identity of *Palaeospondylus gunni* Traquair, a tiny (5–60-mm) vertebrate fossil from the Middle Devonian period (~385 Myr ago) of Scotland, first discovered in 1890 (refs 1–3). It is known principally from a single site (Achanarras Quarry, Caithness) where, paradoxically, it is extremely abundant, preserved in varved lacustrine deposits along with 13 other genera of fishes⁴. Here we show that *Palaeospondylus* is the larval stage of a lungfish, most probably *Dipterus valenciennesi* Sedgwick and Murchison 1828 (ref. 5), and that development of the adult form requires a distinct metamorphosis. *Palaeospondylus* is the oldest known true larva of a vertebrate.

Palaeospondylus gunni has inconclusively been identified in the past as an agnathan, a placoderm, a herring, an amphibian tadpole, a lungfish, a chimaeroid and a shark. Sollas and Sollas made the first full description of this enigmatic fossil, using reconstructed thin sections, but were unable to identify any structures; assuming that *Palaeospondylus* was an adult form with a bony skeleton, they gave each cranial element a special Greek name⁶. However, no histological details are preserved in the coalified tissues; several specimens show traces of connective tissues and some show a distended abdomen, suggesting a yolk-sac (Fig. 1a). It is now clear that the head skeleton consisted largely of endochondral elements in the cartilaginous phase of development, with only a partial dermal covering. Interpreting *Palaeospondylus* as a larva removes many of the difficulties that previous workers encountered in discovering its identity⁷.

Study of whole specimens and new serial-sectioned material (see Supplementary Information) has allowed us to interpret the homology of key cranial elements. The front of the head is dominated by what we interpret as a large attachment organ or sucker, similar to that found in larval *Lepisosteus*⁸. This is supported by the 'rostralia', consisting of two anterior prolongations of the trabeculae cranii and two shorter projections from a single, median, element ('ampyx') identified here as the fused premaxillae. Dorsally, the soft tissue between the dorsal rostralia is preserved in some specimens, showing that it was reinforced by four finger-like extensions from a proximal cartilaginous plate.

The nasal capsules ('hemidomes') are large (Fig. 1b, c) with a delicate, perforated roof, as in modern larval lungfishes (Dipnoi)⁹. There is no sign of lateral narial openings, which were therefore ventrally directed. At least one specimen (Fig. 1d) shows a pair of skeletal rods lying obliquely under the nasal sacs. Various embryonic fishes, including Dipnoi, have subnasal cartilages in this position, but differently oriented¹⁰. It is possible that these structures in *Palaeospondylus* are connected with musculature associated with the attachment organ.

Sollas's 'tauidon' is the vomer, either side of which is a large foramen. Another pair of foramina emerges between the front of the braincase and the premaxillae, possibly for nerves (r. maxillaris V) serving the attachment organ. There is no parasphenoid. The side walls of the anterior braincase form a trough open dorsally, and merge backwards into the sides of the very large otic capsules. Dorsally there are suggestions of a median dermal element (dipnoan B) in the otico-occipital position, and two pairs of dermal elements in the C and D positions.

Teeth or tooth plates are absent, but the 'gammation' and 'posterior trapezoidal bar' are identified here as the palatoquadrate, a large L-shaped structure connected anteriorly to the 'anterior

trapezoidal bar' (probably a rudiment of Meckel's cartilage) and posteriorly to the anterior wall of the otic capsule. The new thin sections show that the posterior skull roof is separated from the endocranium by a pair of spaces suggestive of the adductor cavities of fossil and recent Dipnoi.

Although the hyoid arch (epihyal, ceratohyal and basibranchial) is massive, no other branchial arches are preserved. A characteristic of *Palaeospondylus* is the pair of 'post-occipital lamellae', flattened distally and articulating with a boss on the ventro-lateral corner of the occiput. These are cranial ribs and the orientation of the occipital boss shows that in life they were directed postero-ventrally. Arc-shaped elements generally interpreted as part of the pectoral girdle (Fig. 1c, d) are found closely associated with the cranial ribs, suggesting that, as in living Dipnoi, they were bound to them by musculature in an adaptation for sucking¹¹.

The absence of jaws and most of the dermal skeleton including scales and tooth plates, plus the enlarged cranial ribs, separate occipital arch, rudimentary limb girdles and the presence of an anterior attachment organ, all indicate that *Palaeospondylus* is a larval stage. These features are consistent over the full range of sizes, suggesting an abrupt metamorphosis to the adult after which the attachment organ was lost and a full dermal skeleton developed. All except one of the characteristic features of *Palaeospondylus*—presumed ventral nares, median vomer, dermal skull roof, cranial ribs, ring vertebrae and occipital adductor cavities—are shared with Dipnoi and are absent from the other groups represented in the Caithness faunal assemblage. A dipnoan affinity has been suggested by other authors, such as Kerr, and Forey and Gardiner, but identification of the attachment organ as the nasal capsules prevented a definitive answer to the question^{2,12}. Although an anterior attachment organ is not seen in living lungfishes, *Lepidosiren* and *Protopterus* have an adhesive disc in the gular position. *Palaeospondylus* shows no obvious similarities to the larvae of tetrapods or the adults of arthrodires.

At Achanarras, *Palaeospondylus* occurs principally in an early

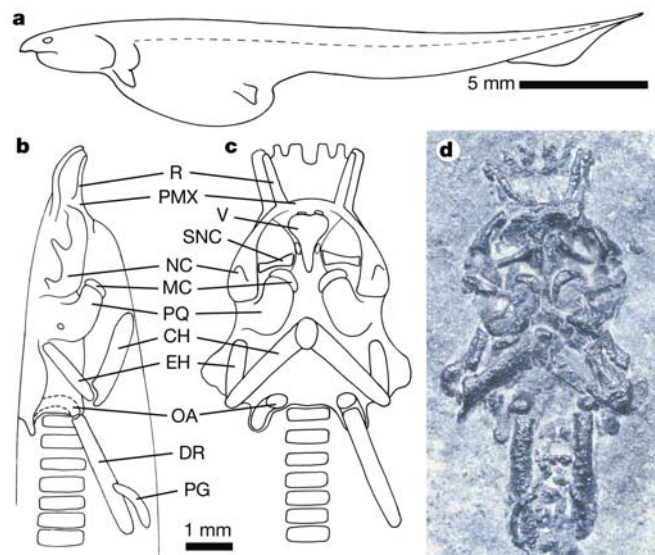


Figure 1 *Palaeospondylus gunni*. **a**, Reconstruction of 25-mm larva with yolk-sac. **b**, Reconstruction of cartilage head skeleton in lateral view. **c**, Reconstruction of cartilage head skeleton in ventral view (right cranial rib removed). **d**, Photograph of head NHM P.59531. EH, epihyal; CH, ceratohyal; DR, cranial rib; MC, Meckel's cartilage; NC, nasal capsule; OA, occipital arch; PG, pectoral girdle; PQ, palatoquadrate; PMX, premaxillae (fused); R, rostralia; SNC, subnasal cartilage; V, vomer.

transgressive phase of lake sedimentation and most closely associated with the equally minute, and equally abundant, acanthodian *Mesacanthus*. It seems likely that they were transported from a sheltered shallow-water environment and deposited *post mortem* in relatively deep water. The exceptional preservation of internal skeletal details might be associated with the absence of squamation¹³. In associated *Mesacanthus* of the same size, the dermal scale covering is complete, making a casing inside which the internal structures would have rotted very quickly before burial.

Dipterus valenciennesi is the only lungfish in the Achanarras fauna and the single most abundant taxon. Therefore, although there are no characters on which a formal synonymy could be based, economy of hypothesis suggests that it is the adult form of *Palaeospondylus gunni*. Evidently this lungfish spent its whole life cycle in fresh water; the early stages involved a separate naked, sessile stage during which it lived off the contents of the yolk-sac and used its attachment organ to attach to vegetation or the substrate. The local abundance and patchy vertical distribution of the larva suggests that the environments in which it lived, before being swept into the deeper water and deposited, were strongly seasonal in occurrence and of limited geographic distribution, such as seasonally flooded, sheltered bays or marshes where the fry of *Mesacanthus* and the larvae of lungfishes could survive. □

Methods

Serial sections were reconstructed to three dimensions by the assembly of images into volume models, the calculation of isosurfaces and the visualization of the resulting surfaces by ray-tracing¹⁴. Images underwent manual 'virtual preparation' before visualization to ensure that the true edge of the fossil was modelled and to apply false colour to discrete structures¹⁵.

Received 20 May; accepted 21 October 2003; doi:10.1038/nature02175.

- Thomson, K. S. The puzzle of *Palaeospondylus*. *Am. Sci.* **80**, 216–219 (1992).
- Forey, P. L. & Gardiner, B. G. J. A. Moy-Thomas and his association with the British Museum (Natural History). *Bull. Br. Mus. Nat. Hist. (Geol.)* **35**, 131–144 (1981).
- Traquair, R. H. On the fossil fishes at Achanarras Quarry, Caithness. *Annu. Mag. Nat. Hist.* **6**(6), 479–486 (1890).
- Trewin, N. Palaeoecology and sedimentology of the Achanarras fish bed of the Middle Old Red Sandstone, Scotland. *Trans. R. Soc. Edinb.* **77**, 21–46 (1986).
- Sedgwick, A. & Murchison, R. I. On the structure and relations of the deposits contained between the Primary Rocks and the Oolitic Series in the North of Scotland. *Trans. Geol. Soc. Lond.* **3**, 125–160 (1828).
- Sollas, W. J. & Sollas, I. B. J. An account of the Devonian fish *Palaeospondylus gunni* Traquair. *Phil. Trans. R. Soc. Lond. B* **196**, 267–294 (1903).
- Moy-Thomas, J. A. The Devonian fish *Palaeospondylus gunni* Traquair. *Phil. Trans. R. Soc. Lond. B* **230**, 391–413 (1940).
- Agassiz, A. The development of *Lepisosteus*. *Proc. Am. Acad. Arts Sci.* **14**, 65–76 (1879).
- Thomson, K. S. The nasal apparatus in Dipnoi, with special reference to *Protopterus*. *Proc. Zool. Soc. Lond.* **145**, 207–238 (1965).
- Fox, H. Early development of the head and pharynx of *Neoceratodus* with a consideration of its phylogeny. *J. Zool.* **146**, 470–554 (1965).
- Bemis, W. E. Feeding systems in living Dipnoi: anatomy and function. *J. Morph. Suppl.* **1**, 249–275 (1987).
- Kerr, J. G. The zoological position of *Palaeospondylus* (Traquair). *Proc. Camb. Phil. Soc.* **10**, 298–299 (1900).
- Rayner, D. H. The Achanarras Limestone of the Middle Old Red Sandstone, Caithness, Scotland. *Proc. Yorks. Geol. Soc.* **34**, 117–138 (1963).
- Sutton, M. D., Briggs, D. E. G., Siveter, David, J. & Siveter, Derek, J. Methodologies for the visualization and reconstruction of three-dimensional fossils from the Silurian Herefordshire lagerstätte. *Paleont. Electron.* **4**(1) article 2, (2000).
- Sutton, M. D., Briggs, D. E. G., Siveter David, J. & Siveter, Derek, J. The arthropod *Offacolus kingi* (Chelicerata) from the Silurian of Herefordshire, England: computer based morphological reconstructions and phylogenetic affinities. *Proc. R. Soc. Lond. B* **269**, 1195–1203 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank the Trustees of the Natural History Museum, London, and the staff of the Palaeontology Department for the loan of specimens, and Derek Siveter for assistance with photography.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to K.S.T. (keith.thomson@university-museum.oxford.ac.uk).

Impact of localized badger culling on tuberculosis incidence in British cattle

Christl A. Donnelly^{1,2}, Rosie Woodroffe^{1,3}, D. R. Cox^{1,4}, John Bourne¹, George Gettinby^{1,5}, Andrea M. Le Fevre², John P. McInerney^{1,6} & W. Ivan Morrison^{1,7}

¹Independent Scientific Group on Cattle TB, c/o Department for Environment, Food & Rural Affairs, 1A Page Street, London SW1P 4PQ, UK

²Department of Infectious Disease Epidemiology, Faculty of Medicine, Imperial College London, St Mary's Campus, Norfolk Place, London W2 1PG, UK

³Department of Wildlife, Fish & Conservation Biology, University of California, One Shields Avenue, Davis, California 95616, USA

⁴Nuffield College, New Road, Oxford OX1 1NF, UK

⁵Department of Statistics and Modelling Science, University of Strathclyde, George Street, Glasgow G1 1XH, UK

⁶Centre for Rural Research, University of Exeter, Lafrowda House, St German's Road, Exeter EX4 6TL, UK

⁷The Centre for Tropical Veterinary Medicine, Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Roslin, Midlothian EH25 9RG, UK

Pathogens that are transmitted between wildlife, livestock and humans present major challenges for the protection of human and animal health, the economic sustainability of agriculture, and the conservation of wildlife. *Mycobacterium bovis*, the aetiological agent of bovine tuberculosis (TB), is one such pathogen. The incidence of TB in cattle has increased substantially in parts of Great Britain in the past two decades, adversely affecting the livelihoods of cattle farmers and potentially increasing the risks of human exposure. The control of bovine TB in Great Britain is complicated by the involvement of wildlife, particularly badgers (*Meles meles*), which appear to sustain endemic infection and can transmit TB to cattle¹. Between 1975 and 1997 over 20,000 badgers were culled as part of British TB control policy, generating conflict between conservation and farming interest groups². Here we present results from a large-scale field trial^{3–5} that indicate that localized badger culling not only fails to control but also seems to increase TB incidence in cattle.

Bovine TB is primarily a disease of cattle but it can be transmitted to people. A 1934 inquiry reported that 40% of dairy cattle in the United Kingdom were infected⁶, with bovine TB causing about 2,000 human deaths annually (6% of the total deaths from TB at that time)⁷. The risk of human infection has been reduced greatly since then, both by pasteurization of milk and by regular tuberculin testing of cattle, with compulsory slaughter of animals showing evidence of exposure. Several countries have effectively controlled TB by this approach, but control has been more difficult to achieve where wildlife act as reservoir hosts that continually re-infect cattle populations. In the British Isles badgers are strongly implicated in transmitting TB to cattle, and badger culling has formed a component of British TB control policy since 1973 (refs 1, 2).

Badger culling strategies have been modified several times^{1,2} (Fig. 1). Badgers were first linked to cattle TB in 1973 and farmers were licensed to cull them. From 1975 to 1981 badgers were gassed in their setts using hydrogen cyanide. This was replaced by the 'clean ring' strategy, designed to identify and remove clusters of infected badgers. From 1986 to 1998 culling occurred only on land used by tuberculin-positive cattle. This was intended as an 'interim strategy' pending development of a more selective culling policy⁸. The effectiveness of these strategies at controlling cattle TB remained unknown because they were not compared with one another, or with a strategy of no culling, in a rigorous randomized trial. The Independent Scientific Group on Cattle TB designed and is overseeing a large-scale trial aimed at evaluating two options of badger

Table 1 History of badger culling in trial areas

Triplet	Treatment	First reactive cull	Badgers culled in reactive treatment	First interim cull	Last interim cull	Years of interim culling	Badgers culled under interim strategy
A	Reactive	2000	117	1992	1997	6	300
	No culling	—	—	1988	1998	7	186
B	Reactive	1999	301	1987	1998	10	314
	No culling	—	—	1987	1997	10	342
C	Reactive	2000	394	1988	1998	7	168
	No culling	—	—	1992	1997	6	319
D	Reactive	2003	122	1995	1997	3	64
	No culling	—	—	1997	1997	1	14
E	Reactive	2002	169	1986	1997	12	455
	No culling	—	—	1986	1998	10	239
F	Reactive	2002	435	1986	1997	10	357
	No culling	—	—	1989	1998	9	240
G	Reactive	2002	256	—	—	0	0
	No culling	—	—	—	—	0	0
H	Reactive	2002	159	1993	1998	5	126
	No culling	—	—	1992	1997	5	31
I	Reactive	2003	94	1997	1998	2	35
	No culling	—	—	1994	1998	5	38
J	Reactive	—	0	1996	1997	2	94
	No culling	—	—	—	—	0	0

The data include reactive culling (in the course of this study) and culling under the interim strategy that preceded this study (1986–98). The fourth column gives the number of badgers culled up to 17 September 2003; the seventh column refers to the number of calendar years in which at least one interim cull occurred.

culling as a means to reduce TB incidence in cattle^{1–5}.

The trial commenced in 1998 and has involved three experimental treatments: (1) localized reactive culling, which seeks to remove badgers from small areas in response to TB outbreaks in cattle; (2) proactive culling, which aims to reduce badger densities to low levels across entire trial areas; and (3) no culling. Of these treatments reactive culling was most similar to past policies, in that it used cattle TB outbreaks as a sentinel for the presence of potentially infectious badgers, whose removal might be expected to reduce the risk of future outbreaks. All aspects of the trial design have been published^{3–5}, and many have been subjected to independent audit^{9–12}. Regular field surveys have indicated changes in badger activity consistent with the aims of the culling treatments; badger activity was lowest in areas subjected to proactive culling, and highest in no culling areas¹³. Trial areas varied in the extent to which they had experienced badger culling in the past (Table 1).

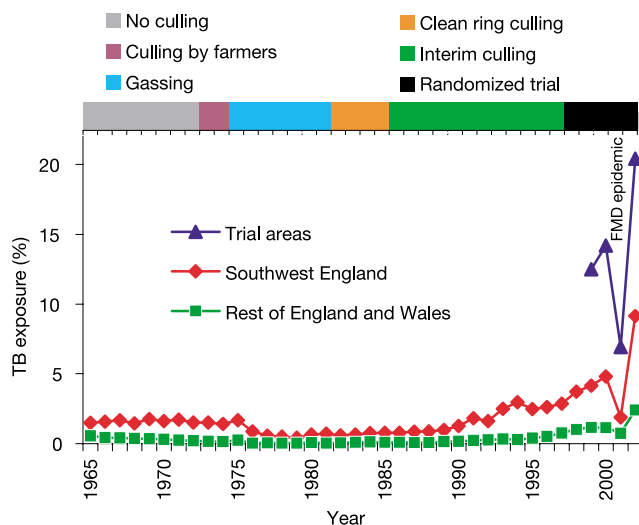


Figure 1 Percentage of total cattle herds found to show evidence of exposure to TB (in confirmed and unconfirmed breakdowns) based on routine testing, and the badger culling policies in operation at the time. Incidence was higher in trial areas because they were placed in areas of highest risk; however, the herds in trial areas experienced temporal trends in TB incidence similar to those throughout England and Wales. Most tests scheduled to occur during the foot-and-mouth disease (FMD) epidemic were delayed until 2002, hence incidence seems low in 2001 and high in 2002.

We compared TB incidence rates (up to 31 August 2003) in cattle herds subjected to the different treatments using log-linear Poisson regression for the number of confirmed breakdowns recorded in each trial area since enrolment. A confirmed herd breakdown is an occasion on which at least one member of a cattle herd showed evidence of TB exposure that was subsequently confirmed by culture of *M. bovis* or by identification of visible lesions of TB at post-mortem examination. In the absence of such confirmation, the breakdown is 'unconfirmed'. A herd breakdown is also confirmed when suspect TB lesions are found in carcasses of animals during routine slaughterhouse meat inspection and those lesions yield *M. bovis* on bacteriological culture. The results presented here concern the effect of reactive culling only.

Our analysis revealed that the reactive treatment has thus far been associated with a 27% increase in the incidence of cattle herd breakdowns ($P = 0.0145$; standard 95% confidence interval of 4.8–53% increase) when compared with no culling areas. This result was highly consistent, with more breakdowns than expected in all nine of the reactively culled areas (Fig. 2). Triplet J had not yet received any reactive culling (Table 1). After adjustment for covariates (see Methods) there was relatively little extra-Poisson variation. The inflation factor (the square root of the model deviance divided by the degrees of freedom) was equal to 1.37 with the goodness-of-

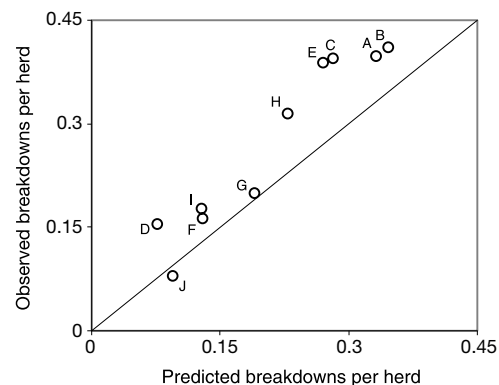


Figure 2 Triplet-specific TB incidence in reactive trial areas (the number of confirmed breakdowns since enrolment in the trial, on completion of the initial proactive cull, divided by the number of baseline herds at risk) observed and predicted had these areas received no culling. Points above the line indicate that increased incidence occurred in reactively culled trial areas. Points have differing precisions and no reactive culling had been carried out in triplet J by September 2003.

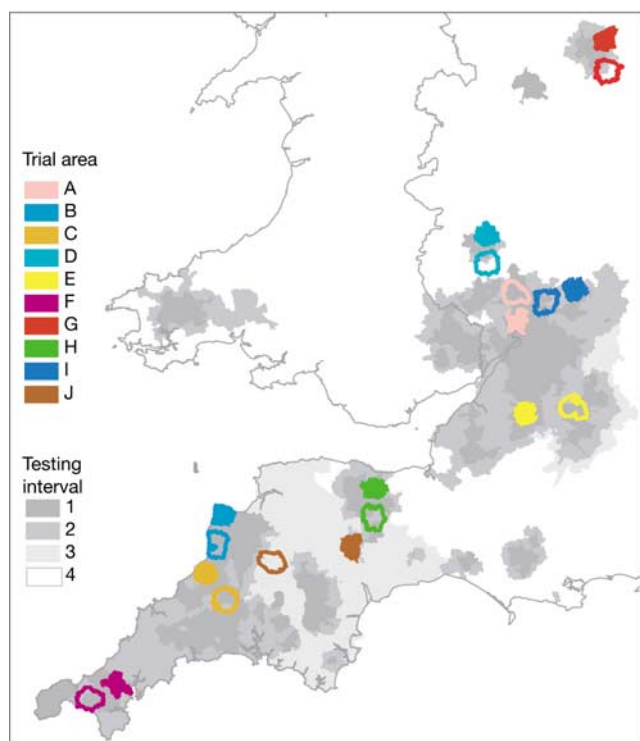


Figure 3 Map of reactive (filled shapes) and no culling (outlined shapes) trial areas superimposed over the 1998 testing intervals for cattle (1-yr testing is conducted in areas of highest incidence, 4-yr testing in areas of lowest incidence). All herds in trial areas have been subjected to annual testing throughout the trial.

fit P -value equal to 0.02. However, even if the confidence interval were conservatively expanded using this factor, the resulting interval—2.4% decrease to 65% increase—does not change the conclusion that localized reactive culling is highly unlikely to decrease substantially TB incidence in cattle.

On the basis of these findings we conclude that reactive culling of badgers, as carried out in the trial, is unlikely to contribute to the control of cattle TB. Our original trial design called for at least one complete year of enrolment in all trial areas (which would occur in late December 2003)⁴; however, as our results show that reactive culling is unlikely to offer a viable base for future policy, we support the decision to discontinue it as an experimental treatment. We recommend, however, that the proactive and no culling treatments be continued until the estimate of any effect is sufficiently precise to inform policy.

Our finding of a consistent increase in cattle TB incidence in response to reactive badger culling provides evidence for a link between badgers and TB in cattle, but points to transmission dynamics that must be highly complex. Control is known to disrupt badger social organization, prompting long-distance movement and dispersal^{14–16}. Such movements are associated with increased TB transmission among badgers¹⁷, and may also influence infection risk for cattle. Studies are underway to investigate this issue.

Our findings may also influence the interpretation of experiments carried out in the Republic of Ireland, where badger culling has been linked with substantial reductions in the incidence of cattle TB¹⁸. Past¹⁸ and ongoing¹⁹ studies compare the incidence of cattle TB in 'project' areas subject to intensive, widespread proactive culling of badgers, with that in 'reference' areas where badgers are culled only in response to TB outbreaks in cattle. As our study demonstrates that reactive culling is associated with an increased incidence of cattle TB relative to no culling, the Irish studies (which lack no-culling experimental control areas) may over-estimate the true effectiveness of widespread proactive culling.

The reactive culling treatment was similar to badger culling policies operating nationwide in Great Britain before the start of the trial. Hence, it is possible that these past approaches may have been similarly ineffective. During the operation of the 'interim' culling strategy, the herd incidence of cattle TB in southwest Britain rose from 0.75% in 1986 to 2.61% in 1996 (ref. 1). Although it is unlikely that this historic increase was entirely attributable to the badger culling policy (Northern Ireland experienced a similar increase in the absence of badger culling²⁰), our results suggest that this form of culling may have been of no benefit to the control of TB in British cattle. □

Methods

Trial design

Thirty trial areas, each covering approximately 100 km², were selected in areas of highest TB risk to cattle (Fig. 3), and were recruited sequentially as ten matched 'triplets'. All trial areas were surveyed for signs of badger activity, and then randomly assigned to treatments in such a way that each treatment was replicated ten times, once within each triplet. Immediately after randomization, an initial proactive cull was carried out in the appropriate area of each triplet. The first such cull was completed in November 1998 and the tenth and final cull in December 2002. Once the initial proactive cull was complete, data on cattle TB incidence in all three trial areas in the triplet were collected, and breakdowns that occurred in reactive areas triggered reactive culling. Reactive culls commenced in nine of the ten triplets thereafter (Table 1). Between March and December 2001 during an outbreak of foot-and-mouth disease (FMD) in Great Britain cattle testing was substantially reduced. Badger culling was suspended at the same time, and recommenced in May 2002.

Reactive culling sought to identify and remove all badgers using the same land as cattle herds that had experienced recent TB outbreaks. Survey data on the distribution of badger dens (setts) and other field signs were used to map approximate home ranges; this method⁸ has been shown to provide a good estimate of the disposition of badger territories when based on good survey data^{10,21}. Culling then took place at all active badger setts falling within home ranges that overlapped the land used by the cattle herd that had shown confirmed evidence of TB exposure. Where the spatial organization of badgers could not be discerned, culling took place at all setts within 1 km of the breakdown farm. When breakdowns occurred on more than one farm in an area, a single culling operation could cover multiple breakdowns. The average area covered by reactive culls (5.3 km²) was intermediate between the extent of historic interim (approx 1 km²) and clean ring (approximately 9 km²) operations¹.

Badgers were captured in cage traps, which were placed primarily at badger setts and baited with peanuts. Badgers were despatched by gunshot, a method deemed "humane" by independent audit^{11,12}. Animals other than badgers were released. All badger carcasses were subjected to full post-mortem examination. No culling took place between 1 February and 30 April each year to avoid killing mothers with unweaned cubs.

Statistical analysis

Data on TB incidence in cattle were analysed by log-linear Poisson regression. The regression model adjusted for triplet, the log of the number of baseline herds at risk, and the log of the number of historic confirmed breakdowns. The historic incidence was calculated for the 3-yr period before the initial cull in the proactive area, except in triplets D, I and J where it was calculated for the 3-yr period before the start of the FMD epidemic.

We tested for covariate interactions with treatment effects for: (1) log of the number of baseline herds at risk; (2) log of the number of historic confirmed breakdowns; (3) per cent of occupiers that agreed to culling; (4) log of the number of active main setts; (5) log of the number of badgers culled during the interim strategy; (6) log of the number of badgers taken in initial and first follow-up proactive culls; (7) TB prevalence in badgers culled in the initial proactive culls; and (8) time since enrolment in the trial. None of these were found to be significant. Thus, there was no evidence that reactive culling performed differently in particular settings.

Furthermore, we analysed alternative response variables for different time periods (pre-FMD, post-FMD and from the end of the first follow-up cull in the proactive area) and for the overall number of breakdowns (including both confirmed and unconfirmed). Models were also fitted replacing the log of the number of baseline herds at risk with either the log of the estimated number of cattle at baseline, the log of the number of tests conducted, or the log of the estimated number of cattle tested. The results obtained were similar in each case.

Received 3 November; accepted 10 November 2003; doi:10.1038/nature02192.

Published online 23 November 2003.

1. Krebs, J. *et al.* *Bovine Tuberculosis in Cattle and Badgers* PB3423 (MAFF Publications, London, 1997).
2. Krebs, J. *et al.* Badgers and bovine TB: Conflicts between conservation and health. *Science* **279**, 817–818 (1998).
3. Bourne, J. *et al.* *Towards a Sustainable Policy to Control TB in Cattle* PB3881 (First Annual Report of the Independent Scientific Group on Cattle TB, MAFF Publications, London, 1998).
4. Bourne, J. *et al.* *An Epidemiological Investigation into Bovine Tuberculosis* PB4870 (Second Annual Report of the Independent Scientific Group on Cattle TB, MAFF Publications, London, 1999).
5. Bourne, J. *et al.* *An Epidemiological Investigation into Bovine Tuberculosis* PB5801 (Third Annual Report of the Independent Scientific Group on Cattle TB, Defra Publications, London, 2001).
6. Hopkins, F. G. *Committee on Cattle Diseases: Report* (HMSO, London, 1934).

7. Hardie, R. M. & Watson, J. M. *Mycobacterium bovis* in England and Wales: past, present and future. *Epidemiol. Infect.* **109**, 23–33 (1992).
8. Woodroffe, R., Frost, S. D. W. & Clifton-Hadley, R. Attempts to control tuberculosis in cattle by removing infected badgers: constraints imposed by live test sensitivity. *J. Appl. Ecol.* **36**, 494–501 (1999).
9. Mollison, D. *TB in Cattle: First Report of the Statistical Auditor on the Badger Culling Trial* PB5385 (MAFF Publications, London, 2000).
10. Cresswell, W. *Report of the Independent Audit of Surveying and Badger Social Group Territory Delineation Procedures in the Randomised Badger Culling Trial and MAFF's Response* PB5497 (MAFF Publications, London, 2001).
11. Kirkwood, J. K. *Report of the Independent Auditor on the Humaneness of the Despatch Procedures Used in the Randomised Badger Culling Trial and MAFF's Response* PB5325 (MAFF Publications, London, 2000).
12. Ewbank, R. *Report of the Second Independent Auditor on the Humaneness of Despatch Procedures Used in the Randomised Badger Culling Trial and Defra's Response* PB8253 (Defra Publications, London, 2003).
13. Le Fevre, A. M. et al. *Changes in Badger Setts Over the First Three Years of the Randomised Badger Culling Trial* (http://www.svepm.org.uk/Posters2003/poster_files/LeFevre.pdf) (2003).
14. Roper, T. J. & Lüps, P. Disruption of territorial behaviour in badgers *Meles meles*. *Z. Säugetierkunde* **58**, 252–255 (1993).
15. Cheeseman, C. L., Cresswell, W. J., Harris, S. & Mallinson, P. J. Comparison of dispersal and other movements in two badger (*Meles meles*) populations. *Mamm. Rev.* **18**, 51–59 (1988).
16. Tuytens, F. A. M., Macdonald, D. W., Rogers, L. M., Cheeseman, C. L. & Roddam, A. W. Comparative study on the consequences of culling badgers (*Meles meles*) on biometrics, population dynamics and movement. *J. Anim. Ecol.* **69**, 567–580 (2000).
17. Rogers, L. M. et al. Movement of badgers (*Meles meles*) in a high-density population: Individual, population and disease effects. *Proc. R. Soc. Lond. B* **265**, 1269–1276 (1998).
18. Eves, J. A. Impact of badger removal on bovine tuberculosis in east County Offaly. *Ir. Vet. J.* **52**, 199–203 (1999).
19. Griffin, J. M. in *Selected Papers 1996* 6–9 (Tuberculosis Investigation Unit, University College Dublin, 1996).
20. King, E. J., Lovell, D. J. & Harris, S. in *Advances in Vertebrate Pest Management* (eds Cowan, D. P. & Feare, C. J.) 147–161 (Filander, Fürth, 1999).
21. Doncaster, C. P. & Woodroffe, R. Den site can determine shape and size of badger territories: implications for group-living. *Oikos* **66**, 88–93 (1993).

Acknowledgements This study was funded and implemented by the Department of Environment, Food and Rural Affairs (DEFRA). We acknowledge the contribution made by staff of DEFRA and its associated agencies. We also wish to thank the many farmers and landowners in the trial areas who allowed the experimental treatments to operate on their land. W. T. Johnston helped prepare Fig. 3.

Authors' contributions J.B., C.A.D., D.R.C., G.G., J.P.M., W.L.M. and R.W. constitute the Independent Scientific Group on Cattle TB, and were jointly responsible for designing and overseeing the study. Statistical analyses were carried out by D.R.C., C.A.D. and A.M.L.F. C.A.D. and R.W. drafted the manuscript, although all authors contributed to its preparation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.A.D. (c.donnelly@imperial.ac.uk).

Predicting distributions of known and unknown reptile species in Madagascar

Christopher J. Raxworthy¹, Enrique Martinez-Meyer², Ned Horning¹, Ronald A. Nussbaum³, Gregory E. Schneider³, Miguel A. Ortega-Huerta² & A. Townsend Peterson⁴

¹American Museum of Natural History, Central Park West at 79th Street, New York, New York 10024-5192, USA

²Instituto de Biología, Universidad Nacional Autónoma de México, Ciudad Universitaria, Mexico City 04510, Mexico

³Museum of Zoology, University of Michigan, Ann Arbor, Michigan 48109-1079, USA

⁴Natural History Museum & Biodiversity Research Center, The University of Kansas, Lawrence, Kansas 66045-2454, USA

Despite the importance of tropical biodiversity¹, informative species distributional data are seldom available for biogeographical study or setting conservation priorities^{2,3}. Modelling ecological niche distributions of species offers a potential solution^{4–7}; however, the utility of old locality data from museums, and

of more recent remotely sensed satellite data, remains poorly explored, especially for rapidly changing tropical landscapes. Using 29 modern data sets of environmental land coverage and 621 chameleon occurrence localities from Madagascar (historical and recent), here we demonstrate a significant ability of our niche models in predicting species distribution. At 11 recently inventoried sites, highest predictive success (85.1%) was obtained for models based only on modern occurrence data (74.7% and 82.8% predictive success, respectively, for pre-1978 and all data combined). Notably, these models also identified three intersecting areas of over-prediction that recently yielded seven chameleon species new to science. We conclude that ecological niche modelling using recent locality records and readily available environmental coverage data provides informative biogeographical data for poorly known tropical landscapes, and offers innovative potential for the discovery of unknown distributional areas and unknown species.

The biota of Madagascar represents a global priority for conservation owing to the island's exceptional endemic diversity and ongoing loss of natural habitats¹. However, although substantial expansion of the national protected area network is anticipated⁸, current rates of deforestation are rapidly reducing the country's future conservation options^{9–11}; in some regions, urgent management decisions will have to be made before detailed biological survey data become available. As a result, the ability to predict biodiversity distribution in poorly known regions of the island offers enormous potential for conservation planning. Using an evolutionary computing approach, we present predictive distribution results for 11 chameleon species, in the first application of satellite imagery to ecological distributional modelling for Madagascar.

For an initial test, with 29 environmental GIS (Geographic Information System) base layers of land cover, climate, topography and hydrology, we used random partitions of our recent post-1988 occurrence localities for building and testing ecological niche models for 11 chameleon species. Using 50% of the post-1988 occurrences to predict the remainder of the post-1988 data, this test yielded impressive results: predictions were significantly better than random for all 11 species under the 'all models predict' criteria and for 9 of the 11 species under the 'any model predicts' criteria, with an overall correct prediction success of 62.8% and 83.0% respectively (Table 1).

A second set of tests included pre-1978 localities^{12,13} for testing models. Models based on the same random 50% of the post-1988 occurrences as described above (for predicting post-1988) were used to predict pre-1978 occurrences. These tests produced significant results for fewer species: 4 and 5 of the 9 species, and overall prediction success was reduced to 33.1% and 63.8% (all models predict and any model predicts criteria, respectively), the reduced prediction ability probably reflecting changes in land use across Madagascar, where recent landscape data (used for modelling) less accurately reflect landscape conditions at the historical time of collecting. To investigate the influence of data density, we repeated this test with models based on 100% of the post-1988 data, and overall results were quite similar or slightly improved (models significant for 3 and 6 of the 9 species, 39.2% and 78.5% overall prediction success under the all models predict and any model predicts assumptions, *respectively). The final temporal partition examined the utility of pre-1978 occurrence data to predict post-1988 distributions, in spite of overall low sample sizes for model building. Predictions resulted in significantly better than random for 6 and 8 of the 9 species, although overall correct prediction success was lower than the other tests (28.0% and 57.1% under the all models predict and any model predicts criteria, respectively).

A third set of tests used data from pre-1978, post-1988 and both temporal partitions combined, for building ecological niche models that were then tested with a completely independent test data set,

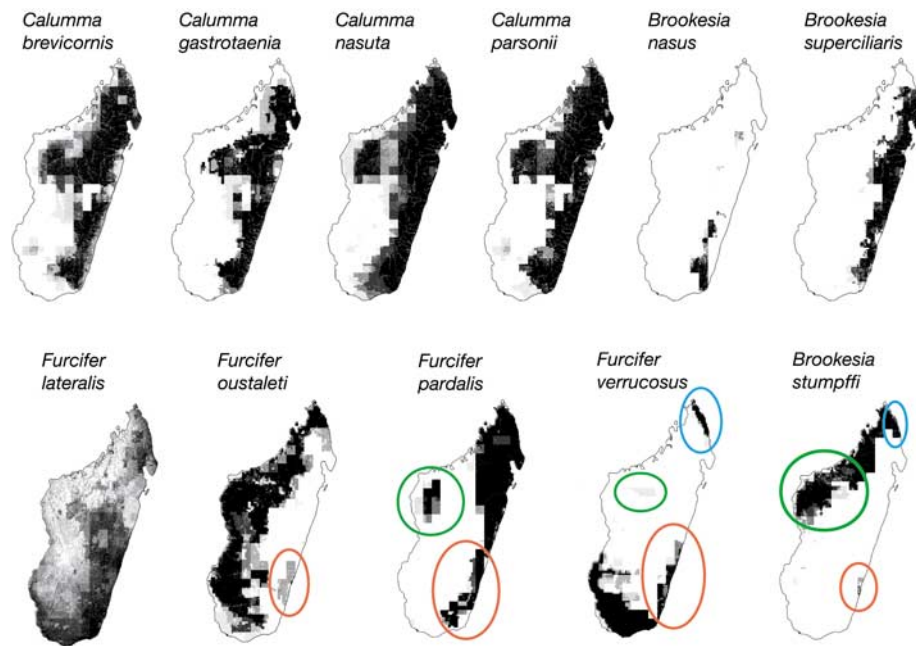


Figure 1 Ecological niche distribution models for 11 chameleon species in Madagascar. The sum of ten best-subset models is shown, with darker shading representing greater model agreement. Areas of over-prediction are circled. The intersecting over-prediction regions in the west (green) and northeast (blue) have recently yielded seven new

locally endemic chameleon species; the other area in the southeast (red) remains poorly studied. The models are based on combined specimen locality data for all species except *F. verrucosus* and *F. oustaleti* (post-1988 locality data only; see Methods).

Table 1 Tests of ecological niche model predictivity

Chameleon taxon	Model training data (N)	Proportion of Madagascar's area predicted by models	Number of correct predictions	Statistical significance
50% post-1988 predicts other 50% post-1988				
<i>Calumma brevicornis</i>	14	0.12–0.30	6 and 10 of 13	*/*
<i>Calumma gastrotaenia</i>	17	0.15–0.29	12 and 16 of 16	*/*
<i>Calumma nasuta</i>	27	0.40–0.63	22 and 27 of 27	*/*
<i>Calumma parsonii</i>	14	0.09–0.32	5 and 10 of 14	*/*
<i>Furcifer lateralis</i>	42	0.52–0.77	28 and 33 of 41	*/—
<i>Furcifer pardalis</i>	19	0.15–0.30	9 and 17 of 19	*/*
<i>Furcifer oustaleti</i>	26	0.39–0.69	15 and 21 of 26	*/—
<i>Furcifer verrucosus</i>	34	0.16–0.28	26 and 29 of 34	*/*
<i>Brookesia nasus</i>	11	0.01–0.03	6 and 10 of 11	*/*
<i>Brookesia stumpffi</i>	11	0.01–0.07	6 and 7 of 11	*/*
<i>Brookesia superciliaris</i>	11	0.06–0.17	5 and 5 of 11	*/*
50% post-1988 predicts all pre-1978				
<i>Calumma brevicornis</i>	14	0.12–0.30	2 and 9 of 14	—/*
<i>Calumma gastrotaenia</i>	17	0.15–0.29	1 and 8 of 11	—/*
<i>Calumma nasuta</i>	27	0.40–0.63	13 and 16 of 21	*/—
<i>Calumma parsonii</i>	14	0.09–0.32	2 and 9 of 20	—/*
<i>Furcifer lateralis</i>	42	0.52–0.77	13 and 21 of 25	—/*
<i>Furcifer pardalis</i>	19	0.15–0.30	4 and 7 of 12	*/*
<i>Brookesia nasus</i>	11	0.01–0.03	2 and 3 of 4	*/*
<i>Brookesia stumpffi</i>	11	0.01–0.07	2 and 2 of 8	*/—
<i>Brookesia superciliaris</i>	11	0.06–0.17	4 and 8 of 15	*/*
All post-1988 predicts all pre-1978				
<i>Calumma brevicornis</i>	27	0.18–0.28	5 and 9 of 14	—/*
<i>Calumma gastrotaenia</i>	33	0.19–0.40	2 and 7 of 11	—/*
<i>Calumma nasuta</i>	54	0.34–0.63	13 and 16 of 21	*/—
<i>Calumma parsonii</i>	28	0.23–0.49	7 and 16 of 20	—/*
<i>Furcifer lateralis</i>	83	0.39–0.95	9 and 24 of 25	—/*
<i>Furcifer pardalis</i>	38	0.21–0.38	5 and 10 of 12	*/*
<i>Brookesia nasus</i>	22	0.02–0.05	2 and 3 of 4	*/*
<i>Brookesia stumpffi</i>	22	0.07–0.13	2 and 4 of 8	—/*
<i>Brookesia superciliaris</i>	22	0.13–0.28	6 and 13 of 15	*/*
All pre-1978 predicts all post-1988				
<i>Calumma brevicornis</i>	14	0.18–0.29	7 and 16 of 27	—/*
<i>Calumma gastrotaenia</i>	11	0.10–0.32	11 and 13 of 33	*/*
<i>Calumma nasuta</i>	21	0.19–0.22	19 and 29 of 54	*/*
<i>Calumma parsonii</i>	20	0.21–0.41	7 and 23 of 28	*/*
<i>Furcifer lateralis</i>	25	0.34–0.76	25 and 57 of 83	—/*
<i>Furcifer pardalis</i>	12	0.08–0.18	10 and 14 of 38	*/*
<i>Brookesia nasus</i>	4	0.01–0.03	3 and 12 of 22	*/*
<i>Brookesia stumpffi</i>	8	0.06–0.13	1 and 8 of 22	—/*
<i>Brookesia superciliaris</i>	15	0.13–0.21	9 and 16 of 22	*/*

Models are based on chameleon locality data from pre-1978 or post-1988, the former tested with data from post-1988 and the latter tested with data from both time periods. Pairs of numbers or symbols refer to statistics calculated for the 'all models predict' and 'any model predicts' criteria, respectively. *, significant prediction ($P < 0.05$); —, not significant.

compiled from 11 other recent site inventories (42 chameleon occurrence localities; see Methods). Here, our evaluation of predictive performance was based on correct predictions of both presence and absence related to total predictions (Table 2): predictive success was 74.7% for pre-1978 models, 85.1% for post-1988 models and 82.8% for combined data models, with uniformly high levels of statistical significance ($P = 3.3 \times 10^{-6}$ for pre-1978, $P = 1.3 \times 10^{-11}$ for post-1988, and $P = 2.8 \times 10^{-10}$ for combined data models).

On the basis of these validation exercises, we conclude that ecological niche modelling of diverse natural history museum data can provide an excellent predictive understanding of reptile species' distributions, even in a region as poorly known as Madagascar. Model predictivity was better when using modern data to predict other modern data, than when using or predicting pre-1978 (historical) data. However, between temporal partitions, predictions significantly better than random were recorded for all species except *Furcifer lateralis*, a widely distributed species for which statistical power was low. Most tellingly, modern and historical museum occurrence information had significant predictive power in indicating where species would and would not be found during subsequent, intensive field studies. Thus, although modern data provide the best predictivity for modelling current species distribution, when modern locality data are rare or exhibit poor spatial coverage, historical data can also provide an excellent guide for understanding the distribution of biodiversity.

Model predictions, when viewed in geographical space (Fig. 1), corresponded closely with our general understanding of chameleon distributions, except that (1) primary humid forest specialists (*Calumma*, *Brookesia*) should be confined to surviving forest fragments within distributional areas, and (2) models for *Furcifer pardalis*, *F. verrucosus*, *F. oustaleti* and *Brookesia stumpffi* show three overlapping predicted areas outside known distributional areas. Because humid forest fragments can be identified using aerial photography or satellite imagery^{9–11} (in addition to verification on the ground), the first exception has little impact on conservation planning; instead, for deforested areas, these models provide a novel perspective concerning the reduction in forest-specialist species' distributions as a consequence of human colonization and environmental degradation. This is particularly the case for the now almost completely deforested Central High Plateau. For example, all four *Calumma* species show a marked and previously unknown distribution limit between the northern and southern parts of the plateau (approximately 19°S), which challenges the generally accepted assumption of a single central ecoregion biota⁸.

Our previous work has documented the conservative nature of ecological niche evolution, which provides predictivity of distributions for closely related species^{14,15}. For closely related allopatric species pairs, niche conservatism is manifested as areas of over-prediction in each species model. Of the three intersecting areas of over-prediction mentioned above (Fig. 1), the southeastern area (between Mahanoro, Vangaindrano and Beraketa) remains largely

Table 2 Distribution of chameleons at 11 independent test sites

Inventory site and model data	<i>C. brevicornis</i> (14, 27, 41)	<i>C. gastrotaenia</i> (11, 33, 44)	<i>C. nasuta</i> (21, 54, 75)	<i>C. parsonii</i> (20, 28, 48)	<i>F. lateralis</i> (25, 83, 108)	<i>F. pardalis</i> (12, 38, 50)	<i>F. oustaleti</i> (52)	<i>F. verrucosus</i> (68)	<i>B. nasus</i> (4, 22, 26)	<i>B. stumpffi</i> (8, 22, 30)	<i>B. supercilialis</i> (15, 22, 37)
Berara						*	*			*	
Pre-1978	–	–	–	–	–	+	NA	NA	–	+	–
Post-1988	–	–	+	–	+	+	+	–	–	+	–
Combined	–	–	+	–	+	+	NA	NA	–	+	–
NE Manongarivo	*	*				*				*	
Pre-1978	–	–	–	–	+	–	NA	NA	–	+	–
Post-1988	+	–	+	+	+	–	–	–	–	+	–
Combined	+	–	+	+	–	+	NA	NA	–	+	–
Bemanevika	*	*	*		*						
Pre-1978	–	+	–	–	+	–	NA	NA	–	–	–
Post-1988	+	–	+	+	+	–	–	–	–	–	–
Combined	+	+	+	+	+	+	NA	NA	–	+	–
Lohanandroranga	*	*	*								
Pre-1978	–	+	–	–	+	–	NA	NA	–	–	–
Post-1988	+	+	+	+	+	–	–	–	–	–	–
Combined	+	+	+	+	–	+	NA	NA	–	–	–
Tampolo			*								*
Pre-1978	+	–	–	–	–	+	NA	NA	–	–	–
Post-1988	–	–	+	–	–	–	–	–	–	–	–
Combined	–	–	+	–	–	+	NA	NA	–	–	–
Andranomay	*	*	*	*	*						
Pre-1978	+	+	+	+	+	–	NA	NA	–	–	+
Post-1988	+	+	+	+	–	–	–	–	–	–	+
Combined	+	+	+	+	+	–	NA	NA	–	–	+
Ambatovy	*	*	*								*
Pre-1978	+	+	+	+	+	–	NA	NA	–	–	+
Post-1988	+	+	+	–	–	–	–	–	–	–	+
Combined	+	+	+	+	–	+	NA	NA	–	–	+
Ankazomivady	*	*	*		*						
Pre-1978	+	+	+	–	–	–	NA	NA	–	–	–
Post-1988	+	–	–	–	–	–	–	–	–	–	–
Combined	+	+	+	+	+	–	NA	NA	–	–	–
And./Ran. corridor	*	*	*	*	*				*		*
Pre-1978	+	+	+	+	+	–	NA	NA	+	–	+
Post-1988	+	+	+	+	+	–	–	–	+	–	+
Combined	+	+	+	+	+	–	NA	NA	+	–	+
Ivohibe corridor	*	*	*	*					*		*
Pre-1978	–	–	–	+	+	–	NA	NA	–	–	–
Post-1988	+	+	+	+	–	–	–	–	+	–	+
Combined	+	+	+	+	–	–	NA	NA	+	–	+
Bemanteza								*			
Pre-1978	–	–	–	–	+	–	NA	NA	–	–	–
Post-1988	–	–	–	–	–	–	–	+	–	–	–
Combined	–	–	–	–	–	–	NA	NA	–	–	–

The 11 sites were recently inventoried, with predictions (presence and absence) based on ecological niche models (pre-1978, post-1988, or combined occurrence data, and the 'any model predicts' criteria). Model sample sizes (pre-1978, post-1988, combined) are given in parentheses for each species. *, recorded present during inventory; +, predicted present; –, predicted absent; NA, not applicable (see Methods); And./Ran., Andringitra–Ranomafana.

unexplored, but we have recently surveyed the two other areas. Although neither area was previously considered rich in locally endemic *Furcifer* or *Brookesia* chameleons^{12,13}, the western area (northern Bongolava Mountains, between Soalala, Maintirano and Tsiroanomandidy) yielded an undescribed *Furcifer* and an undescribed *Brookesia* species. The other area, in extreme northeastern Madagascar (transitional humid/dry forests between Daraina, Vohemar and Sambava), yielded five undescribed *Brookesia* and a possible new *Furcifer* species. These discoveries, of at least seven new species, sharply contrast with our results elsewhere in Madagascar: more intensive surveys at sites outside these areas of over-prediction over the same time period have produced only two other undescribed *Furcifer* and *Brookesia* chameleons. Hence, ecological niche modelling not only provides powerful predictions of species' distributions, but also offers an innovative and predictive approach to discovering heretofore unknown populations of known or unknown species.

These findings have valuable applications for identification of areas of endemism³, and consequently will also aid the development of inclusive strategies for conserving regional endemism¹⁶. In Madagascar, such a model-based evaluation of species' distributions in the context of existing reserves and protected corridors, coupled with surveys targeted on areas of predicted potentially unrecognized endemism or distribution, will provide data critical for effective future expansion of the protected area network⁸. Our results also highlight the continued importance of obtaining modern locality data, because of their optimal predictive performance when used with remotely sensed data. Notably, because equivalent landscape coverage data and specimen locality data are available for most other regions worldwide, this approach offers the potential for providing informative distribution data, and directing surveys to unknown populations or species, for many other poorly known and threatened environments. □

Methods

Species occurrence data

Three suites of species occurrence information were used in our analyses: pre-1978 occurrences drawn from natural history museum specimens^{12,13} (130 unique locality × species combinations; *F. verrucosus* and *F. oustaleti* excluded due to potential taxonomic confusion), post-1988 data obtained from 1989–2002 surveys by C.J.R. and R.A.N. (449 unique locality × species combinations), and independent test occurrence data obtained from other recent herpetological inventories^{17–24} (42 unique locality × species combinations). Post-1988 data are vouched for by specimens deposited at the American Museum of Natural History and the University of Michigan Museum of Zoology. Pre-1978 locality descriptions were converted to geographical coordinates using 1:100,000 topographic maps and gazetteers.

All Madagascar chameleon species or species/subspecies complexes²⁵ for which ≥20 occurrence localities were available from post-1988 were included: *Calumma gastrotaenia* (includes subspecies as recognized by refs 11 and 12), *C. nasuta* (*C. fallax*), *C. parsonii* (*C. oshaughnessyi*, *C. globifer*) and *Brookesia superciliaris* (*B. thezezi*), where potential synonyms are indicated in parentheses. Test data in the form of additional herpetological inventories met the following criteria: rainy season surveys (November to April) made by experienced herpetologists, designed to sample complete chameleon diversity, at sites ≥7 km from occurrence localities used in modelling. Meeting these criteria were 11 sites that included diverse primary habitats (humid to dry forests), elevation of 0–1,700 m and latitude of 14–23° S: Berara¹⁷; northeast Manongarivo²⁰; Tampolo²⁴; Andranomay²²; Ankazomivady¹⁸; Andringitra–Ranomafana corridor²¹; Ivohibe corridor²³; Bemananteza¹⁹; and unpublished inventory data for Bemanevika Lakes (14° 20' S, 48° 35' E), Lohanandroranga (14° 25' S, 49° 09' E) and Ambatovy (18° 51' S, 48° 19' E) (Raxworthy/University of Antananarivo).

Ecological and environmental coverage data

Initially, 29 GIS coverages were considered for inclusion in models. A land cover layer was derived from 17 MODIS 1-km resolution, 16-d composite, BRDF/albedo-adjusted reflectance images (MOD43B4) from November 2000 to November 2001. Unsupervised classification methods created a layer with 50 classes, subsequently grouped and edited manually into 13 land cover classes using higher resolution imagery for reference. To summarize precipitation, we used 5' resolution data (1996–2001) from the NOAA Climate Prediction Center Famine Early Warning System (FEWS) (<http://edcw2ks21.cr.usgs.gov/adds/>): annual minimum, maximum and mean precipitation estimates were taken from 10-d cumulative estimates provided by FEWS. Data on cloud cover, mean temperature in January, July and annually, and precipitation in January, July and annually were obtained from ESRI²⁶; maximum, minimum, mean temperatures and mean diurnal temperature range for dry season, wet season and annually for 1961–1990

(0.5° resolution) were obtained from ref. 27 (<http://ipcc-ddc.cru.uea.ac.uk/>). Finally, topographical data (elevation, slope, aspect, flow accumulation, flow direction and tendency to pool water; 1 km resolution) were obtained from the US Geological Survey (<http://edcdaac.usgs.gov/gtopo30/hydro/>). All data and coverages were re-sampled to 0.01° (about 1 km) resolution. Sets of geographical coverages were reduced to an optimal 25 ecological dimensions via a process of jack-knifing inclusion of dimensions in analyses and inspection of omission statistics²⁸.

Ecological niche modelling

We used the Genetic Algorithm for Rule-set Prediction (GARP)^{4,5}, an evolutionary computing system that has excellent capacities for delineating ecological niches and predicting geographical distributions of species^{6,29}. GARP niche models outline potential ecological distributions, and as such roughly map onto Hutchinson's fundamental niche concept (as opposed to the realized niche space, which takes into account the limiting effects of interactions with other species)²⁹. GARP models were developed using a PC desktop implementation of the program (<http://www.lifemapper.org/desktopgarp/>).

GARP produces sets of rules delineating ecological niches by relating known occurrence points to a suite of GIS data layers describing the ecological landscape. Input points are separated into training and testing samples to permit model refinement based on independent samples of points. Simple inferential tools produce initial rules (roughly in the form of linear combinations of coverages), which serve as seeds for search and improvement in the genetic algorithm. Rules 'evolve' through an iterative process of random rule selection, evaluation, perturbation, testing, and incorporation or rejection, and the evolutionary process stops at convergence^{4,5}. For each taxon, we produced 100 replicate models and a best subset of 10 models was selected based on optimal combinations of error components²⁹; models that predicted <90% of testing points were discarded, and (of the remaining models) the ten closest to the median predicted area were summed to provide a best distributional prediction.

Predictive abilities of models were evaluated using pre-1978, post-1988 and combined occurrence data, testing models with independent suites of points not included in model building (Table 1). Because sample sizes for each species were generally small, we used binomial tests to evaluate statistical significance of predictions³⁰. Observed successes and failures were compared with random expectations based on the proportional area predicted present by the intersection of all ten or any of the ten best-subset models. For validation of models using the independent inventory data described above, a cross-species test using a χ^2 2 × 2 contingency table (1 degree of freedom) was applied.

Received 27 August; accepted 11 November 2003; doi:10.1038/nature02205.

- Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B. & Kent, J. Biodiversity hotspots for conservation priorities. *Nature* **403**, 853–858 (2000).
- Pimm, S. L., Russell, G. J., Gittleman, J. L. & Brooks, T. M. The future of biodiversity. *Science* **269**, 347–350 (1995).
- Crisci, J. V. The voice of historical biogeography. *J. Biogeogr.* **28**, 157–168 (2001).
- Stockwell, D. R. B. & Noble, I. R. Induction of sets of rules from animal distribution data: A robust and informative method of analysis. *Math. Comput. Simul.* **33**, 385–390 (1992).
- Stockwell, D. R. B. & Peters, D. P. The GARP modelling system: Problems and solutions to automated spatial prediction. *Int. J. Geogr. Info. Syst.* **13**, 143–158 (1999).
- Peterson, A. T., Stockwell, D. R. B. & Klzua, D. A. in *Predicting Species Occurrences: Issues of Accuracy and Scale* (eds Scott, J. M., Heglund, P. J. & Morrison, M. L.) 617–623 (Island Press, Washington DC, 2002).
- Peterson, A. T. *et al.* Future projections for Mexican faunas under global climate change scenarios. *Nature* **416**, 626–629 (2002).
- ANGAP. *Plan de Gestion du Réseau National des Aires Protégées de Madagascar* (Association National pour la Gestion des Aires Protégées, Antananarivo, Madagascar, 2001).
- Green, G. M. & Sussman, R. W. Deforestation history of the eastern rain forests of Madagascar from satellite images. *Science* **248**, 212–215 (1990).
- Nelson, R. & Horning, N. AVHRR-LAC estimates of forest area in Madagascar, 1990. *Int. J. Remote Sens.* **14**, 1463–1475 (1993).
- Mayaux, P., Gond, V. & Bartholomé, E. Mapping the forest-cover of Madagascar with SPOT 4-VEGETATION data. *Int. J. Remote Sens.* **21**, 3139–3144 (2000).
- Brygoo, E. R. Reptiles Sauriens Chamaeleontidae. Genre *Chamaeleo*. *Faune Madagascar* **33**, 1–318 (1971).
- Brygoo, E. R. Reptiles Sauriens Chamaeleontidae. Genre *Brookesia* et complément pour le genre *Chamaeleo*. *Faune Madagascar* **47**, 1–174 (1978).
- Peterson, A. T., Soberón, J. & Sánchez-Cordero, V. Conservatism of ecological niches in evolutionary time. *Science* **285**, 1265–1267 (1999).
- Martinez-Meyer, E. *Evolutionary Trends in Ecological Niches of Species*. Thesis, Univ. Kansas (2002).
- Mittermeier, R. A., Myers, N., Thomsen, J. B. & da Fonseca, G. A. B. Biodiversity hotspots and major tropical wilderness areas: Approaches to setting conservation priorities. *Conserv. Biol.* **12**, 516–520 (1998).
- Andreone, F., Vences, M. & Randrianirina, J. E. Patterns of amphibian and reptile diversity at Berera Forest (Sahamalaza Peninsula), NW Madagascar. *Ital. J. Zool. (Modena)* **68**, 235–241 (2001).
- Goodman, S. M. *et al.* Inventaire biologique de la Forêt d'Ankazomivady, Ambositra. *Akon'ny Ala* **24**, 19–32 (1998).
- Goodman, S. M. *et al.* Inventaire des vertébrés du Parc National de Tsimanampetsotsy (Toliara). *Akon'ny Ala* **28**, 1–36 (2002).
- Rakotomalala, D. Diversité des reptiles et amphibiens de la Réserve Spéciale de Manongarivo, Madagascar. *Boissiera* **59**, 339–358 (2002).
- Rakotomalala, D., Raholimavo, E., Talata, P. & Rajeriarison, E. Les amphibiens et reptiles du Parc National de Ranomafana et de la zone forestière le reliant au Parc National d'Andringitra. *Rech. Dév. Série Sci. Biol.* **17**, 133–163 (2001).
- Raselimanana, A. P. Inventaire biologique, Forêt d'Andranomay, Anjozorobe: La diversité de la faune de reptiles et d'amphibiens. *Rech. Dév. Série Sci. Biol.* **13**, 43–59 (1998).
- Raselimanana, A. P. Inventaire biologique de la Réserve Spéciale de Pic d'Ivohibe et du couloir

forestier qui la relie au Parc National d'Andringitra: L'herpetofaune. *Rech. Dév. Série Sci. Biol.* 15, 81–97 (1999).

24. Rasselimanana, A. P., Rakotomalala, D. & Rakotondraparany, F. Inventaire biologique de la forêt littorale de Tampolo (Fenoarivo Atsinanana). Les reptiles et amphibiens: Diversité et conservations. *Rech. Dév. Série Sci. Biol.* 14, 183–195 (1998).
25. Raxworthy, C. J., Forstner, M. R. J. & Nussbaum, R. A. Chameleon radiation by oceanic dispersal. *Nature* 415, 784–787 (2002).
26. ESRI. *ArcAtlas* (Environmental Systems Research Institute, Redlands, California, 1997).
27. New, M., Hulme, M. & Jones, P. A 1961–1990 *Mean Monthly Climatology of Global Land Areas* (Climatic Research Unit, University of East Anglia, Norwich, 1997).
28. Peterson, A. T. & Cohoon, K. C. Sensitivity of distributional prediction algorithms to geographic data completeness. *Ecol. Mod.* 117, 159–164 (1999).
29. Anderson, R. P., Lew, D. & Peterson, A. T. Evaluating predictive models of species' distributions: Criteria for selecting optimal models. *Ecol. Mod.* 162, 211–232 (2003).
30. Anderson, R. P., Laverde, M. & Peterson, A. T. Geographical distributions of spiny pocket mice in South America: Insights from predictive models. *Global Ecol. Biogeogr.* 11, 131–141 (2002).

Acknowledgements We thank the Malagasy authorities and the University of Antananarivo Department of Animal Biology for their assistance. Fieldwork (C.J.R. and R.A.N.) was supported by Conservation International, Earthwatch, the National Geographic Society, the US National Science Foundation and the World Wide Fund for Nature. This work was supported by NASA and the Center for Biodiversity and Conservation at the American Museum of Natural History.

Competing interests statement The authors declare they have no competing financial interests.

Correspondence and requests for materials should be addressed to C.J.R. (rax@amnh.org).

Presynaptic induction of heterosynaptic associative plasticity in the mammalian brain

Yann Humeau¹, Hamdy Shaban¹, Stephanie Bissière¹ & Andreas Lüthi^{1,2}

¹Friedrich Miescher Institute, Maulbeerstrasse 66, CH-4058 Basel, Switzerland

²Department of Pharmacology/Neurobiology, Biozentrum, University of Basel, Klingelbergstrasse 70, CH-4056 Basel, Switzerland

The induction of associative synaptic plasticity in the mammalian central nervous system classically depends on coincident presynaptic and postsynaptic activity^{1,2}. According to this principle, associative homosynaptic long-term potentiation (LTP) of excitatory synaptic transmission can be induced only if synaptic release occurs during postsynaptic depolarization^{1,2}. In contrast, heterosynaptic plasticity in mammals is considered to rely on activity-independent, non-associative processes^{3–8}. Here we describe a novel mechanism underlying the induction of associative LTP in the lateral amygdala (LA). Simultaneous activation of converging cortical and thalamic afferents specifically induced associative, N-methyl-D-aspartate (NMDA)-receptor-dependent LTP at cortical, but not at thalamic, inputs. Surprisingly, the induction of associative LTP at cortical inputs was completely independent of postsynaptic activity, including depolarization, postsynaptic NMDA receptor activation or an increase in postsynaptic Ca²⁺ concentration, and did not require network activity. LTP expression was mediated by a persistent increase in the presynaptic probability of release at cortical afferents. Our study shows the presynaptic induction and expression of heterosynaptic and associative synaptic plasticity on simultaneous activity of converging afferents. Our data indicate that input specificity of associative LTP can be determined exclusively by presynaptic properties.

Bipolar stimulating electrodes were placed on afferent fibres from the internal capsule (containing thalamic afferents)^{9–11} or from the external capsule (containing cortical afferents)¹² in coronal slices prepared from 3–4-week-old male C57BL/6J mice (Fig. 1a). Whole-cell current-clamp recordings were obtained from projection

neurons in the dorsolateral portion of the LA (Fig. 1a). Low-frequency baseline stimulation in the presence of the GABA_A (γ-aminobutyric acid) receptor antagonist picrotoxin (100 μM) elicited monosynaptic excitatory postsynaptic potentials (EPSPs) of similar amplitudes and slopes at both afferent inputs (thalamic, 5.6 ± 0.4 mV, 1.07 ± 0.11 mV ms^{−1}; cortical, 5.7 ± 0.4 mV, 1.04 ± 0.11 mV ms^{−1}; n = 13). To mimic the physiological activity of converging thalamic and cortical afferents during fear conditioning^{13,14}, both afferents were stimulated simultaneously for 1.5 s at an average frequency of 30 Hz using two different stimulation protocols containing Poisson-distributed stimuli ('Poisson-train'; Fig. 1b; see Methods). Simultaneous Poisson-train stimulation resulted in the induction of LTP at cortical (151 ± 10% of baseline, n = 13, P < 0.01), but not at thalamic, afferent synapses (98 ± 5%, n = 13, P > 0.05; Fig. 1c). Inverting the two stimulation patterns to assess stimulation protocol-specific effects did not affect the input-specific induction of LTP at cortical input synapses (cortical, 152 ± 16% of baseline, n = 6, P < 0.05; thalamic, 106 ± 12%, n = 6, P > 0.05). The induction of LTP was associative, in that stimulation of both the thalamic and cortical afferents was required. Stimulation of either pathway on its own did not induce LTP at cortical afferents (cortical, 106 ± 15% of baseline, n = 6, P > 0.05; thalamic, 101 ± 8%, n = 5, P > 0.05; Fig. 1d), or at thalamic afferents (cortical, 100 ± 8% of baseline, n = 7, P > 0.05; thalamic, 105 ± 16%, n = 5, P > 0.05; see Supplementary Information), indicating that the stimulation protocols applied were below threshold for the induction of homosynaptic^{10–12} and heterosynaptic LTP at cortical afferents.

Associative LTP in the hippocampus² and the amygdala^{11,12} depends largely on the activation of NMDA receptors and an increase in the postsynaptic Ca²⁺ concentration. Accordingly, heterosynaptic, associative LTP (LTP_{HA}) at cortical afferents could not be induced in the presence of the competitive NMDA receptor antagonist 3-(±)-2-carboxypiperazin-4-yl-propyl-1-phosphonic acid (CPP) at 20 μM (control, 151 ± 10% of baseline, n = 13; CPP, 88 ± 8% of baseline, n = 9, P > 0.05; Fig. 2a). To assess whether NMDA receptor activation in conjunction with Poisson-train stimulation of cortical afferents was sufficient for the induction of LTP_{HA}, we applied NMDA locally in the vicinity of the projection neuron from which we were recording by using a pressure application system. Whereas puff-application of NMDA in the absence of cortical afferent activity did not result in the induction of LTP_{HA} (98 ± 4% of baseline, n = 5, P > 0.05; Fig. 2b), combining the application of NMDA with Poisson-train stimulation of cortical afferents resulted in a potentiation of cortical afferent synapses (157 ± 12% of baseline, n = 4, P < 0.05; Fig. 2b). In contrast, pairing NMDA application with Poisson-train stimulation of thalamic afferents did not induce LTP at thalamic afferents (99 ± 10% of baseline, n = 3, P > 0.05; Fig. 2b).

To determine whether an increase in postsynaptic Ca²⁺ concentration was required for LTP_{HA} induction we dialysed the postsynaptic neuron with the Ca²⁺ chelator BAPTA (10–50 mM). Surprisingly, postsynaptic dialysis with BAPTA did not prevent the induction of LTP_{HA} (152 ± 17% of baseline, n = 14, P < 0.05; Fig. 2c). Given that activation of NMDA receptors is required for the induction of LTP_{HA}, this finding suggests that they are not located on the postsynaptic neuron or, alternatively, that they can signal in a Ca²⁺-independent way. To test these possibilities we dialysed the postsynaptic cell with the NMDA receptor open-channel blocker MK-801, and stimulated cortical and thalamic afferents while holding the postsynaptic cell at +30 mV (ref. 15). This procedure completely blocked postsynaptic NMDA receptors (Fig. 2d). However, even the complete blockade of postsynaptic NMDA receptors did not interfere with the induction of LTP_{HA} (134 ± 9% of baseline, n = 4, P < 0.05; Fig. 2d). To test whether Ca²⁺ signalling was required, we next incubated the slices with BAPTA-acetoxymethyl ester (BAPTA-AM; 50 μM), a membrane-

permeant form of the Ca^{2+} chelator BAPTA. BAPTA-AM completely abolished LTP_{HA} induction ($86 \pm 8\%$ of baseline, $n = 7$, $P > 0.05$; Fig. 2e). In conclusion, the induction of LTP_{HA} was independent of postsynaptic activity including depolarization, postsynaptic NMDA receptor activation and increase in intracellular Ca^{2+} concentration, but still required NMDA receptor activation and Ca^{2+} signalling.

We considered the possibility that LTP_{HA} induction might represent a network phenomenon involving NMDA receptors located on other neurons within the LA. We therefore sought to decrease network excitability strongly in the LA, a brain structure that is tightly controlled by GABA-mediated inhibition^{16,17}, by application of the GABA_A receptor agonist muscimol (5 μM) during LTP induction. Indeed, heterosynaptic forms of plasticity mediated by network activity have been shown to be strongly reduced by activation of GABA_A receptors^{5,18}. Muscimol clamped the membrane potential at or near the chloride equilibrium potential (Fig. 3a), thereby preventing action potential initiation in all neurons expressing GABA_A receptors, including projection neurons and local inhibitory interneurons^{19–21}. However, LTP_{HA} induction was not affected by the presence of muscimol ($146 \pm 14\%$ of baseline, $n = 6$, $P < 0.05$; Fig. 3b), but we could not exclude the possibility that some unknown factor released by neurons not expressing GABA_A receptors would be required for the induction of LTP_{HA} . To investigate this further, we strongly suppressed network activity by application of the AMPA (α -amino-3-hydroxy-5-methyl-4-isoxazole propionic acid) receptor antagonist NBQX (2,3-dihydroxy-6-nitro-7-sulphamoyl-benzo(f)quinoxaline; 20 μM). However, even when AMPA-receptor-mediated synaptic transmission was completely blocked, LTP_{HA} could still be induced as monitored by NMDA-receptor-mediated excitatory postsynaptic currents (EPSCs) recorded at +30 mV ($122 \pm 8\%$ of baseline, $n = 7$, $P < 0.05$; Fig. 3c). Thus, most parsimoniously, our

data are consistent with the possibility that glutamate released by thalamic afferents might directly activate NMDA receptors located on presynaptic terminals of cortical afferents, a hypothesis supported by electron-microscopic studies indicating the presence of the NMDA receptor subunit NR1 on presynaptic terminals in the LA^{22–24}.

If activation of NMDA receptors on presynaptic terminals of cortical afferents underlies the induction of LTP_{HA} , this raises the question why Poisson-train stimulation of cortical afferents alone does not induce LTP. One possible explanation is that glutamate released at cortical afferents would be rapidly cleared by glutamate uptake. Indeed, we found that a single Poisson-train stimulation of cortical afferents was able to induce LTP in the presence of a low concentration of the glutamate uptake blocker TBOA (D,L-threo- β -benzyloxyaspartate; 20 μM ; $133 \pm 9\%$ of baseline, $n = 5$, $P < 0.05$; Fig. 3d). To assess whether the facilitation of LTP induction at cortical afferents in the presence of TBOA was due to the activation of presynaptic NMDA receptors by increased ambient glutamate levels, we checked whether TBOA affected the presynaptic properties of cortical afferents (see below) or spontaneous excitatory network activity. However, TBOA did not significantly affect paired-pulse facilitation (PPF) at cortical afferents (control, 1.35 ± 0.15 ; TBOA, 1.44 ± 0.22 ; $n = 3$, $P > 0.05$) or the frequency of spontaneous EPSPs (control, 3.0 ± 0.7 Hz; TBOA, 3.2 ± 1.1 Hz; $n = 5$, $P > 0.05$).

Given the induction mechanism of LTP_{HA} , we reasoned that thalamic afferent stimulation should affect presynaptic function of cortical afferents in an NMDA-receptor-dependent manner. We therefore compared PPF²⁵ in response to double stimulation of cortical afferents (inter-stimulus interval 50 ms) before and after tetanic stimulation (45 stimuli at 30 Hz) of thalamic afferents. Tetanic stimulation of thalamic afferents resulted in a transient decrease in PPF at cortical afferents ($82 \pm 6\%$ of baseline, $n = 7$,

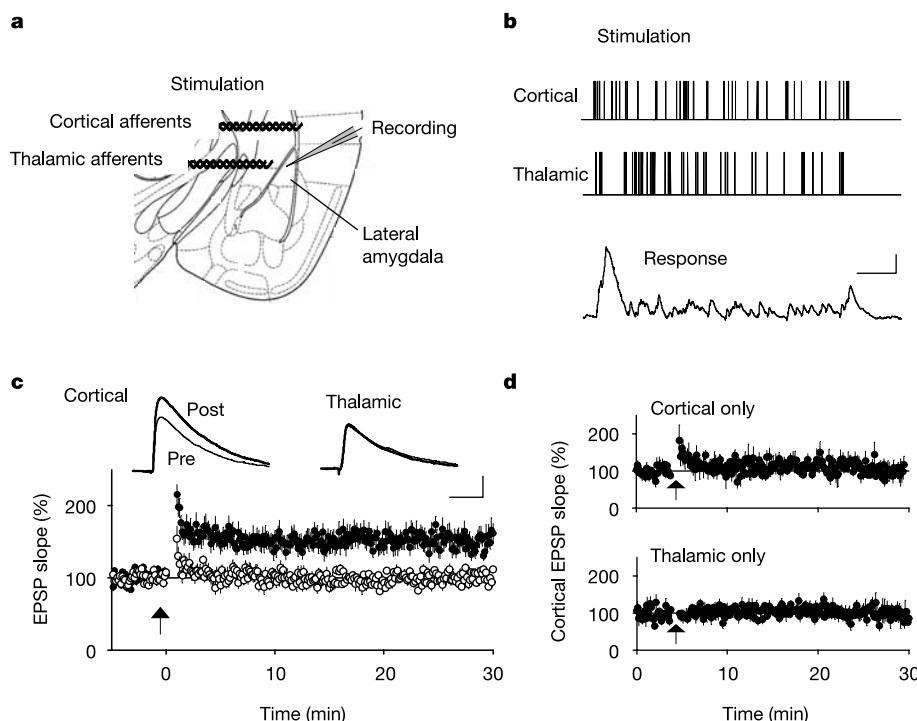


Figure 1 Induction of LTP_{HA} at cortical, but not at thalamic, afferent synapses by simultaneous Poisson-train stimulation of thalamic and cortical afferents. **a**, Placement of stimulating and recording electrodes. **b**, Poisson-train stimulation used for LTP induction. Each train consisted of 45 stimuli at an average frequency of 30 Hz. The trace shows a typical postsynaptic response. Scale bars, 10 mV and 250 ms. **c**, Time course of synaptic

changes after simultaneous Poisson-train stimulation (arrow) of cortical (filled circles) and thalamic (open circles) afferents. Scale bars, 2 mV and 50 ms. **d**, Time course of synaptic changes occurring at cortical afferent synapses upon Poisson-train stimulation (arrow) of either cortical or thalamic afferents alone.

$P < 0.05$; Fig. 4a–c). In contrast, tetanic stimulation of cortical afferents did not affect PPF at thalamic afferents ($104 \pm 4\%$ of baseline, $n = 5$, $P > 0.05$; Fig. 4c). Furthermore, the decrease in cortical PPF after thalamic afferent stimulation was completely

abolished by the NMDA receptor antagonist CPP ($20 \mu\text{M}$; $107 \pm 5\%$ of baseline, $n = 6$, $P > 0.05$; Fig. 4a, c). These results indicate that repeated stimulation of thalamic afferents transiently increases the probability of release (P_r) at cortical afferents in an

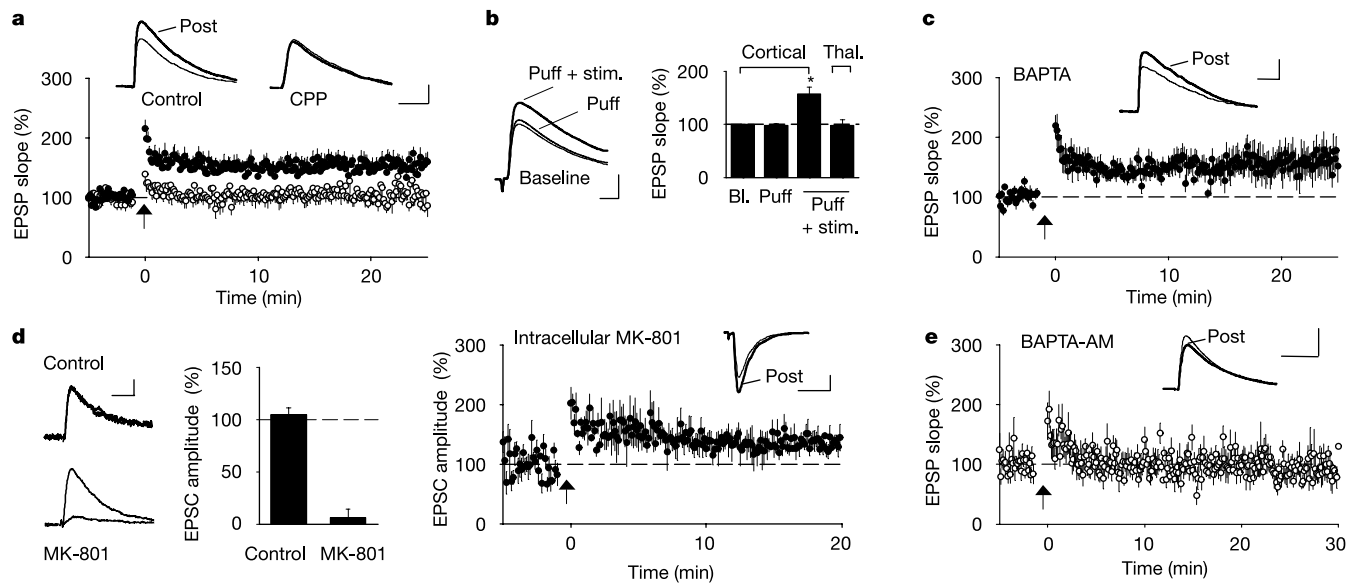


Figure 2 Induction of LTP_{HA} does not depend on postsynaptic activity, but is dependent on NMDA receptor activation and Ca^{2+} . **a**, Induction of LTP_{HA} is blocked in NMDA receptor antagonist CPP (control, same data as in Fig. 1c). Filled circles, control; open circles, CPP. Scale bars, 2 mV and 50 ms. **b**, Left, averaged EPSPs from 2 min of baseline, after pressure application of NMDA (Puff), and 25 min after pairing of NMDA application with cortical afferent stimulation (Puff + stim.). Scale bars, 2 mV and 25 ms. Right, changes in EPSP slope induced by pressure application of NMDA alone, and in conjunction with Poisson-train stimulation of cortical or thalamic afferents. **c**, Induction of LTP_{HA} is

independent of increase in postsynaptic Ca^{2+} . Scale bars, 2 mV and 50 ms. **d**, Left, intracellular dialysis with MK-801 blocks NMDA-receptor-mediated EPSCs recorded at +30 mV in the presence of NBQX. Traces show averaged NMDA EPSCs for the first and last five stimulations. Scale bars, 20 pA and 100 ms. Middle, MK-801-induced blockade of NMDA-receptor-mediated EPSCs at cortical and thalamic afferents (pooled data). Right, LTP_{HA} is not affected by blockade of postsynaptic NMDA receptors. Scale bars, 50 pA and 20 ms. **e**, Induction of LTP_{HA} is blocked in the presence of BAPTA-AM. Scale bars, 2 mV and 50 ms.

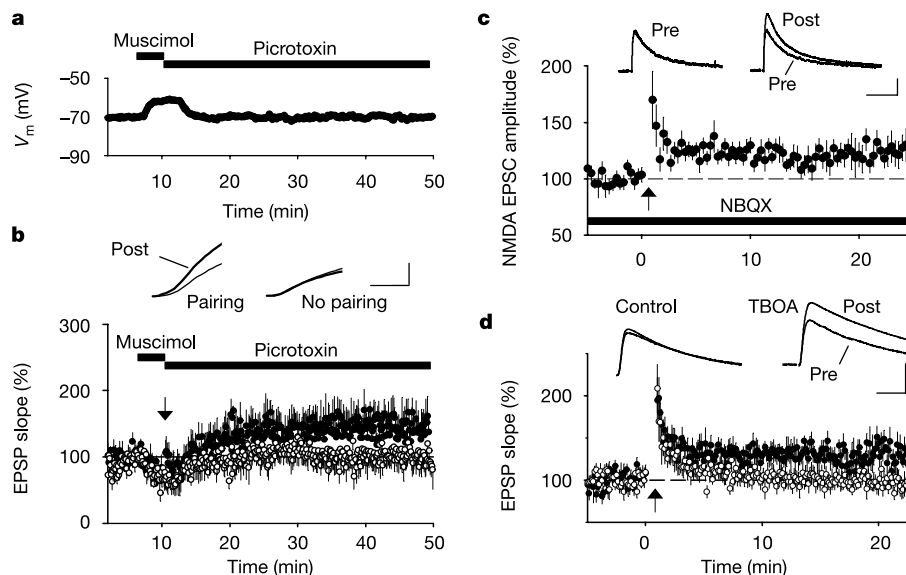


Figure 3 Network activity is not required for the induction of LTP_{HA} . **a**, Average time course of the membrane potential (V_m) during muscimol application. **b**, Induction of LTP_{HA} in the presence of muscimol. Time course of cortical EPSP slope after Poisson-train stimulation and in the absence of Poisson-train stimulation. PicROTOXIN was used to terminate shunting induced by tonic GABA_A receptor activation. Filled circles, pairing; open circles, no pairing. Scale bars, 1.3 mV and 2 ms. **c**, Induction of LTP_{HA} in the presence of NBQX. The plot shows a time course of NMDA-receptor-mediated EPSCs at cortical afferents recorded at +30 mV. Scale bars, 20 pA and 100 ms. **d**, Induction of homosynaptic LTP at cortical afferents in the presence of the glutamate uptake blocker TBOA. Open circles, control; filled circles, TBOA. Scale bars, 3 mV and 25 ms.

open circles, no pairing. Scale bars, 1.3 mV and 2 ms. **c**, Induction of LTP_{HA} in the presence of NBQX. The plot shows a time course of NMDA-receptor-mediated EPSCs at cortical afferents recorded at +30 mV. Scale bars, 20 pA and 100 ms. **d**, Induction of homosynaptic LTP at cortical afferents in the presence of the glutamate uptake blocker TBOA. Open circles, control; filled circles, TBOA. Scale bars, 3 mV and 25 ms.

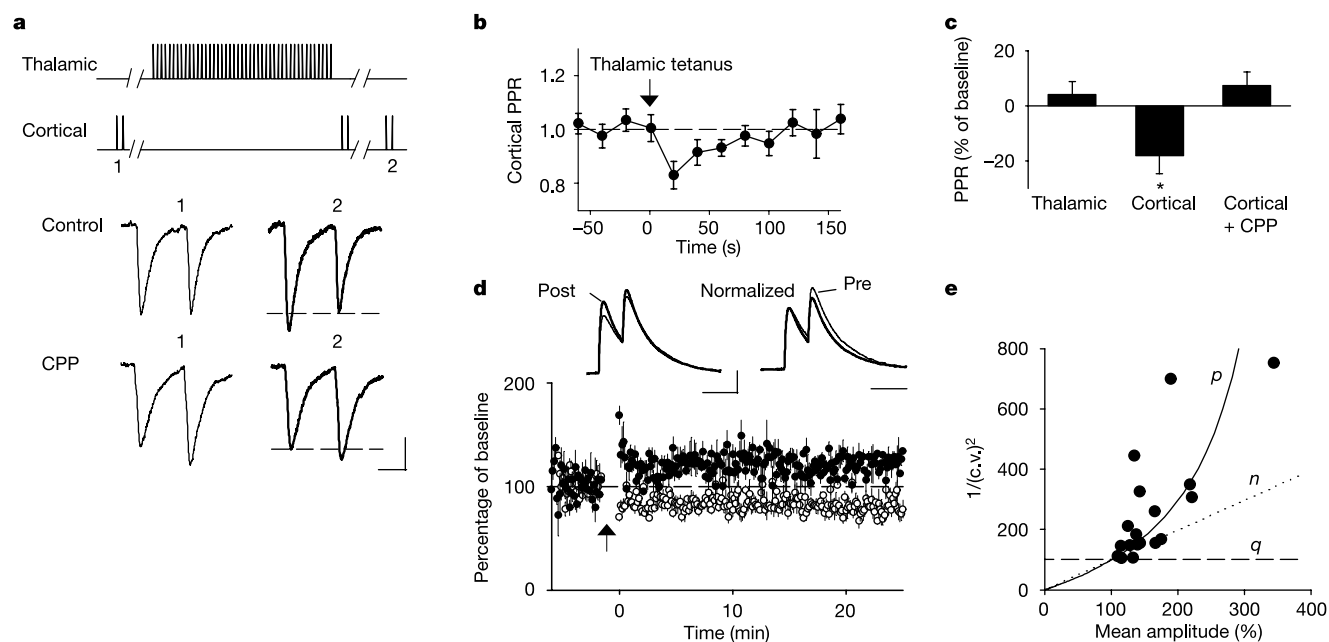


Figure 4 Heterosynaptic, NMDA-receptor-dependent increase of release probability at cortical afferents after stimulation of thalamic afferents. **a**, Top, stimulation protocols. Bottom, EPSCs recorded in response to paired-pulse stimulation (50-ms inter-stimulus interval) of cortical afferents before (point 1) and 10 s after (point 2) tetanic stimulation of thalamic afferents in control conditions or in the presence of CPP. Dashed lines represent the initial amplitude (point 1) of the first response before thalamic stimulation. Scale bars,

40 pA and 20 ms. **b**, Average time course of PPF changes at cortical afferent synapses after tetanic stimulation of thalamic afferents. **c**, PPF changes at thalamic and at cortical afferents after tetanic stimulation of the other input. **d**, Average time course of cortical EPSP slope (filled circles) and PPF (open circles) after induction of LTP_{HA}. Scale bars, 2.5 mV and 75 ms. **e**, Variance analysis of EPSP amplitude fluctuations illustrating that LTP_{HA} expression fits best with an increase in the probability of release (p).

NMDA-receptor-dependent manner and that the expression of LTP_{HA} might involve a more persistent increase in P_r at cortical synapses. Indeed, LTP_{HA} was associated with a persistent decrease in PPF ($81 \pm 5\%$ of baseline, $n = 6$, $P < 0.05$; Fig. 4d). Moreover, if LTP_{HA} expression were mediated by an increase in P_r , we reasoned that it should be occluded at high initial P_r . Therefore, we increased P_r by increasing the extracellular Ca^{2+} concentration from 2 to 8 mM. This induced an increase in EPSP amplitude to $175 \pm 19\%$ ($n = 8$, $P < 0.05$) and occluded the further induction of LTP_{HA} after adjusting EPSP amplitude to control values ($111 \pm 11\%$ of baseline, $n = 5$, $P > 0.05$). Finally, using analysis of fluctuations of the postsynaptic response amplitude, we determined the quantal parameters modified upon induction of LTP_{HA} (refs 26, 27). The plot of $1/(\text{c.v.})^2$ (where c.v. is the coefficient of variation) against mean response amplitude shows that LTP_{HA} induction fits best with an increase in P_r (Fig. 4e). Finally, because $1/(\text{c.v.})^2$ is independent of the quantal amplitude (q), we directly assessed possible changes in q by monitoring the amplitude of asynchronously released quanta in the presence of Sr^{2+} under control conditions and after the induction of LTP_{HA}. The amplitude of stimulation-induced miniature EPSCs at cortical afferents was not significantly affected by the induction of LTP_{HA} (control, -6.7 ± 0.2 pA; LTP, -7.4 ± 1.1 pA; $n = 4$, $P > 0.05$). Because homosynaptic LTP at cortical afferent synapses has previously been shown to be mediated by a presynaptic increase in P_r (refs 12, 26), our results indicate that heterosynaptic LTP_{HA} could share the same expression mechanism. Indeed, we found that the prior induction of homosynaptic LTP by pairing afferent stimulation at 2 Hz with postsynaptic depolarization^{12,26} occluded the subsequent induction of LTP_{HA} using Poisson-train stimulation (see Supplementary Information). However, in contrast to the induction of homosynaptic LTP, which depends on postsynaptic NMDA receptor activation and can be blocked by postsynaptic BAPTA (30 mM; $103 \pm 10\%$ of baseline, $n = 5$)^{12,26},

LTP_{HA} is induced and expressed presynaptically.

The physiological relevance of LTP_{HA} for fear learning is supported by two observations. First, thalamic and cortical afferents to the lateral amygdala are simultaneously active during fear conditioning and can interact^{9,28}. Second, electrophysiological experiments *in vivo* show that postsynaptic hyperpolarization does not completely abolish LTP induced by pairing of sensory stimulation with foot-shocks¹⁴, suggesting that LTP induction independent of postsynaptic activity does occur *in vivo*. The physiological stimulation patterns used in this study suggest that LTP_{HA} might be induced on subthreshold activity elicited by simultaneous sensory input by means of thalamic and cortical afferents. LTP_{HA} might therefore serve as a priming mechanism to increase the impact of selective cortical afferents on the subsequent induction of homosynaptic hebbian plasticity at neighbouring synapses, which requires stronger afferent activity and/or the induction of postsynaptic action potentials^{10–12,20,29,30}.

Our study indicates that the input specificity of associative LTP can be entirely determined by presynaptic properties. Heterosynaptic associative modifications of synaptic efficacy add a level of complexity to the classical hebbian forms of synaptic plasticity, and open a new perspective for understanding integrative processes between converging afferent pathways in the mammalian central nervous system. □

Methods

Coronal slices from 3–4-week-old male C57BL/6J mice were prepared as described²⁰. Slices were maintained for 45 min at 35 °C in an interface chamber containing artificial cerebrospinal fluid equilibrated with 95% O₂/5% CO₂ and containing (in mM): 124 NaCl, 2.7 KCl, 2 CaCl₂, 1.3 MgCl₂, 26 NaHCO₃, 0.4 NaH₂PO₄, 10 glucose, 4 ascorbate; and then kept for at least 45 min at 21–25 °C before being transferred to a superfusing recording chamber. Whole-cell recordings were performed at 30–32 °C. Neurons were identified visually with infrared videomicroscopy. Patch electrodes (3–5 MΩ) were normally filled with a solution containing (in mM): 120 potassium gluconate, 20 KCl, 10 HEPES, 10 phosphocreatine, 4 Mg-ATP, 0.3 Na-GTP, pH 7.25, 295 mOsm. All experiments were

performed in the presence of picrotoxin (100 μ M) unless indicated otherwise. Monosynaptic EPSPs exhibiting constant 10–90% rise times and latencies were elicited by stimulation of afferent fibres with a bipolar twisted platinum/10% iridium wire (25 μ m diameter). LTP was quantified for statistical comparisons by normalizing and averaging EPSP slopes during the last 10 min of experiments relative to 5–10 min of baseline. Depicted traces show averaged EPSPs or EPSCs for 2 min of baseline and 2 min of LTP (20–25 min after pairing). All values are expressed as means $\pm$ s.e.m. Statistical comparisons were done with paired or unpaired Student's *t*-test as appropriate (two-tailed *P* < 0.05 was considered significant). For details on experimental conditions and analysis see Supplementary Information.

Received 18 August; accepted 24 October 2003; doi:10.1038/nature02194.

- Gustafsson, B. & Wigström, H. Basic features of long-term potentiation in the hippocampus. *Semin. Neurosci.* **2**, 321–333 (1990).
- Bliss, T. V. P. & Collingridge, G. L. A synaptic model of memory: long-term potentiation in the hippocampus. *Nature* **361**, 31–39 (1993).
- Bailey, C. H., Giustetto, M., Huang, Y. Y., Hawkins, R. D. & Kandel, E. R. Is heterosynaptic modulation essential for stabilizing Hebbian plasticity and memory? *Nature Rev. Neurosci.* **1**, 11–20 (2000).
- Artola, A. & Singer, W. Long-term depression of excitatory synaptic transmission and its relationship to long-term potentiation. *Trends Neurosci.* **16**, 480–487 (1993).
- Scanziani, M., Malenka, R. C. & Nicoll, R. A. Role of intercellular interactions in heterosynaptic long-term depression. *Nature* **380**, 446–450 (1996).
- Tsukamoto, M. *et al.* Mossy fibre synaptic NMDA receptors trigger non-Hebbian long-term potentiation at entorhino-CA3 synapses in the rat. *J. Physiol. (Lond.)* **546**, 665–675 (2003).
- Royer, S. & Paré, D. Conservation of total synaptic weight through balanced synaptic depression and potentiation. *Nature* **422**, 518–522 (2003).
- Nishiyama, M., Hong, K., Mikoshiba, K., Poo, M. M. & Kato, K. Calcium stores regulate the polarity and input-specificity of synaptic modification. *Nature* **408**, 584–588 (2000).
- LeDoux, J. E. Emotion circuits in the brain. *Annu. Rev. Neurosci.* **23**, 155–184 (2000).
- Weisskopf, M. G., Bauer, E. P. & LeDoux, J. E. L-type voltage gated calcium channels mediate NMDA-independent associative long-term potentiation at thalamic input synapses to the amygdala. *J. Neurosci.* **19**, 10512–10519 (1999).
- Bauer, E. P., Schafe, G. E. & LeDoux, J. E. NMDA receptors and L-type voltage gated calcium channels contribute to long-term potentiation and different components of fear memory formation in the lateral amygdala. *J. Neurosci.* **22**, 5239–5249 (2002).
- Huang, Y. Y. & Kandel, E. R. Postsynaptic induction and PKA-dependent expression of LTP in the lateral amygdala. *Neuron* **21**, 169–178 (1998).
- Quirk, G. J., Armony, J. L. & LeDoux, J. E. Fear conditioning enhances different temporal components of tone-evoked spike trains in auditory cortex and lateral amygdala. *Neuron* **19**, 613–624 (1997).
- Rosenkranz, J. A. & Grace, A. A. Dopamine-mediated modulation of odour-evoked amygdala potentials during pavlovian conditioning. *Nature* **417**, 282–287 (2002).
- Berretta, N. & Jones, R. S. Tonic facilitation of glutamate release by presynaptic N-methyl-D-aspartate autoreceptors in the entorhinal cortex. *Neuroscience* **75**, 339–344 (1996).
- Lang, E. J. & Paré, D. Similar inhibitory processes dominate the responses of cat lateral amygdaloid projection neurons to their various afferents. *J. Neurophysiol.* **77**, 341–352 (1997).
- Szinyei, C., Heinbockel, T., Montagne, J. & Pape, H. C. Putative cortical and thalamic inputs elicit convergent excitation in a population of GABAergic interneurons of the lateral amygdala. *J. Neurosci.* **20**, 8909–8915 (2000).
- Abraham, W. C. & Wickens, J. R. Heterosynaptic long-term depression is facilitated by blockade of inhibition in area CA1 of the hippocampus. *Brain Res.* **546**, 336–340 (1991).
- Lang, E. J. & Paré, D. Synaptic responsiveness of interneurons of the cat lateral amygdaloid nucleus. *Neuroscience* **83**, 877–889 (1998).
- Bissière, S., Humeau, Y. & Lüthi, A. Dopamine gates LTP induction in lateral amygdala by suppressing feedforward inhibition. *Nature Neurosci.* **6**, 587–592 (2003).
- McDonald, A. J. & Mascagni, F. Immunohistochemical localization of the $\beta 2$ and $\beta 3$ subunits of the GABA_A receptor in the basolateral amygdala of the rat and monkey. *Neuroscience* **75**, 407–419 (1996).
- Farb, C. R., Aoki, C. & LeDoux, J. E. Differential localization of NMDA and AMPA receptor subunits in the lateral and basal nuclei of the amygdala: a light and electron microscopic study. *J. Comp. Neurol.* **362**, 86–108 (1995).
- Gracy, K. N. & Pickel, V. M. Ultrastructural localization of NMDAR1 glutamate receptor immunoreactivity in the extended amygdala. *J. Comp. Neurol.* **362**, 71–85 (1995).
- Farb, C. R. & LeDoux, J. E. Afferents from the rat temporal cortex synapse on lateral amygdala neurons that express AMPA and NMDA receptors. *Synapse* **33**, 218–229 (1999).
- Hess, G., Kuhnt, U. & Voronin, L. L. Quantal analysis of paired-pulse facilitation in guinea pig hippocampal slices. *Neurosci. Lett.* **77**, 187–192 (1987).
- Tsvetkov, E., Carlezon, W. A., Benes, F. M., Kandel, E. R. & Bolshakov, V. Y. Fear conditioning occludes LTP-induced presynaptic enhancement of synaptic transmission in the cortical pathway to the lateral amygdala. *Neuron* **34**, 289–300 (2002).
- Humeau, Y., Popoff, M. R., Kojima, H., Doussau, F. & Poulain, B. Rac GTPase plays an essential role in exocytosis by controlling the fusion competence of release sites. *J. Neurosci.* **22**, 7968–7981 (2002).
- Doyère, V., Schafe, G. E., Sigurdsson, T. & LeDoux, J. E. Long-term potentiation in freely moving rats reveals asymmetries in thalamic and cortical inputs to the lateral amygdala. *Eur. J. Neurosci.* **17**, 2703–2715 (2003).
- Markram, H., Lübke, J., Frotscher, M. & Sakman, B. Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science* **275**, 213–215 (1997).
- Blair, H. T., Schafe, G. E., Bauer, E. P., Rodrigues, S. M. & LeDoux, J. E. Synaptic plasticity in the lateral amygdala: a cellular hypothesis of fear conditioning. *Learn. Mem.* **8**, 229–242 (2001).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank B. Gähwiler, C. Heuss, A. Matus, B. Poulain and E. Seifritz for discussions and comments on the manuscript. This work was supported by the Borderline Personality Disorder Research Foundation, the Swiss National Science Foundation and the Novartis Research Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to A.L. (andreas.luthi@fmi.ch).

A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*

Robert J. Johnston Jr & Oliver Hobert

Department of Biochemistry and Molecular Biophysics, Center for Neurobiology and Behavior, Columbia University, College of Physicians and Surgeons, 701 W.168th Street, New York, New York 10032, USA

How left/right functional asymmetry is layered on top of an anatomically symmetrical nervous system is poorly understood. In the nematode *Caenorhabditis elegans*, two morphologically bilateral taste receptor neurons, ASE left (ASEL) and ASE right (ASER), display a left/right asymmetrical expression pattern of putative chemoreceptor genes that correlates with a diversification of chemosensory specificities^{1,2}. Here we show that a previously undefined microRNA termed *lsey-6* controls this neuronal left/right asymmetry of chemosensory receptor expression. *lsey-6* mutants that we retrieved from a genetic screen for defects in neuronal left/right asymmetry display a loss of the ASEL-specific chemoreceptor expression profile with a concomitant gain of the ASER-specific profile. A *lsey-6* reporter gene construct is expressed in less than ten neurons including ASEL, but not ASER. *lsey-6* exerts its effects on ASEL through repression of *cog-1*, an Nkx-type homeobox gene, which contains a *lsey-6* complementary site in its 3' untranslated region and that has been shown to control ASE-specific chemoreceptor expression profiles³. *lsey-6* is the first microRNA to our knowledge with a role in neuronal patterning, providing new insights into left/right axis formation.

Bilateral symmetry is a common feature of nervous system anatomy across phylogeny. Morphological symmetry is contrasted by the lateralization of many different nervous system functions, as well as by left/right asymmetrical patterns of gene expression^{4–6}. However, in only a very few cases have these asymmetrical gene expression patterns been specifically correlated with functional lateralization. One such example is provided by a bilaterally symmetrical class of taste receptor neurons, the ASE class^{7,8}, which displays left/right asymmetrical expression patterns of a family of putative chemoreceptors with guanylyl cyclase activity (Fig. 1a). In adult animals, the guanylyl cyclase receptor genes *gcy-6* and *gcy-7* are only expressed in ASEL, whereas *gcy-5* is only expressed in ASER¹. The asymmetry of *gcy* gene expression patterns correlates with a functional asymmetry of the ASEL and ASER neurons exemplified by their sensation of a different spectrum of chemosensory cues². The differential segregation of the chemosensory capacities in the left and right ASE neurons is required for intact behaviour in complex sensory environments. In mutant animals that exhibit a partial lateralization of chemoreceptor expression, chemosensation *per se* is unaffected, yet the ability to discriminate between individual chemosensory inputs is lost^{2,9}.

To elucidate how left/right asymmetrical expression of the *gcy*

chemoreceptors is established, we conducted a genetic screen for mutants in which the normally ASEL-specific expression of a *gcy-7^{prom}::gfp* reporter gene is disrupted³. One mutant retrieved from this screen, termed *lsy-6* (for laterally symmetrical), displays a 'two ASER' mutant phenotype, in which ASEL fails to show its normal expression of *gcy-7*, but instead expresses the normally ASER-specific *gcy-5* gene (Fig. 1b, c). Other aspects of ASE cell fate specification are unaffected (see below). Left/right asymmetrical olfactory receptor expression in the AWC neurons^{6,10} and directional left/right asymmetrical Q neuroblast migration⁶ is also unaffected in *lsy-6* mutants (data not shown).

Through standard three-factor mapping with a polymorphic *C. elegans* isolate¹¹, we mapped *lsy-6* to a 55-kilobase region between two single-nucleotide polymorphisms on chromosome V (Fig. 2a). After further narrowing the *lsy-6* locus through transformation rescue with DNA fragments from this region, we sequenced the locus in *lsy-6* mutants and uncovered a 1,071-base-pair (bp) deletion. Subfragments of the deleted region rescued the mutant phenotype, yet lacked any predicted protein product (Fig. 2a). Comparing the sequence of the rescuing fragments of *C. elegans* to the syntenic region of *Caenorhabditis briggsae*, a related nematode species, we identified a conserved inverted repeat that has the capacity to form a hairpin structure (Fig. 2b). Hairpin structures of this kind have previously been reported to be processed to 22/23-nucleotide-long microRNAs (miRNAs) that repress messenger RNA translation of specific target genes through binding to their 3' untranslated region (UTR)^{12–16}. Within the *C. elegans* and *C. briggsae* hairpin structures lies a highly conserved 23-nucleotide sequence that probably forms a biologically active miRNA species. Outside of this 23-nucleotide region, the hairpin structures show limited primary sequence conservation but strong conservation of their secondary structures (Fig. 2b). Sequence similarity searches between the *lsy-6* hairpin or miRNA and the entries in the central miRNA registry (<http://www.sanger.ac.uk/Software/Rfam/mirna/search.shtml>) produced no hits, suggesting that *lsy-6* is a novel miRNA, not previously retrieved from either cloning or computer prediction approaches^{14–17}.

A 521-bp fragment containing putative 5' transcriptional regulatory elements of the *lsy-6* locus and the *lsy-6* hairpin structure can

rescue the asymmetry defect, both in terms of regaining *gcy-7* expression in ASEL and losing ectopic *gcy-5* expression in ASEL (Fig. 2a, fragment number 7). When the hairpin sequence is removed from a 781-bp rescuing fragment, rescue is lost (fragment number 9). Conversely, the hairpin sequence alone, without any flanking sequence, is not sufficient to rescue the asymmetry defect (fragment number 10), presumably owing to a lack of transcriptional regulatory elements. A substitution of four nucleotides in the stem of the hairpin abolishes *lsy-6* activity (*mut1*, fragment number 11). Furthermore, a single point mutation (*mut2*) in the hairpin removes the rescuing ability of the 781-bp fragment (number 12). Introduction of a compensatory mutation (*comp mut*) that restores hairpin secondary structure restores the rescuing activity (fragment number 13). Similarly, expression of the complete hairpin in one of the two possible orientations under control of a heterologous promoter is also able to restore ASEL fate (Fig. 3c). These experiments indicate that the hairpin is essential for *lsy-6* activity and that, by inference, *lsy-6* probably encodes a small regulatory RNA.

In order to identify the cellular focus of *lsy-6* action, we made use of the previous observation that miRNA expression can be monitored through RNA polymerase II-dependent transcriptional reporter constructs¹⁸. To this end, two kilobases of the 5' upstream

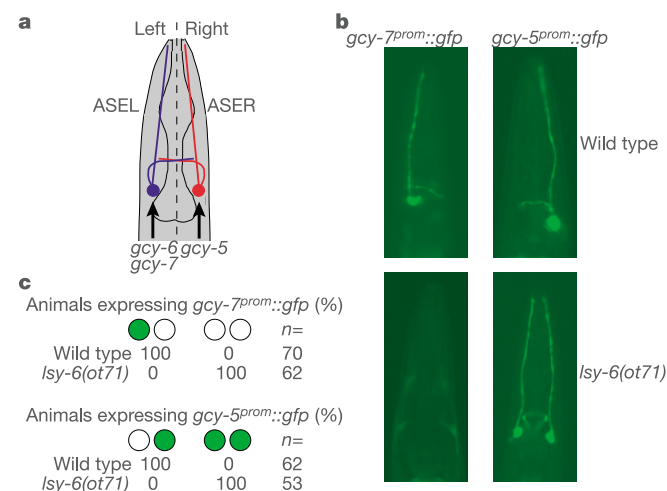


Figure 1 *Lsy-6* regulates asymmetrical chemoreceptor gene expression. **a**, Gene expression asymmetry in the ASE class of chemosensory neurons. **b**, In *lsy-6(ot71)* mutants, *gcy-7^{prom}::gfp* expression is lost and *gcy-5^{prom}::gfp* expression is derepressed in ASEL. **c**, Quantification of *lsy-6(ot71)* effects on asymmetrical *gcy* expression. Circles represent ASEL and ASER, respectively, expressing either one of the indicated chromosomally integrated *gfp* reporter gene arrays. Animals were scored as adults.

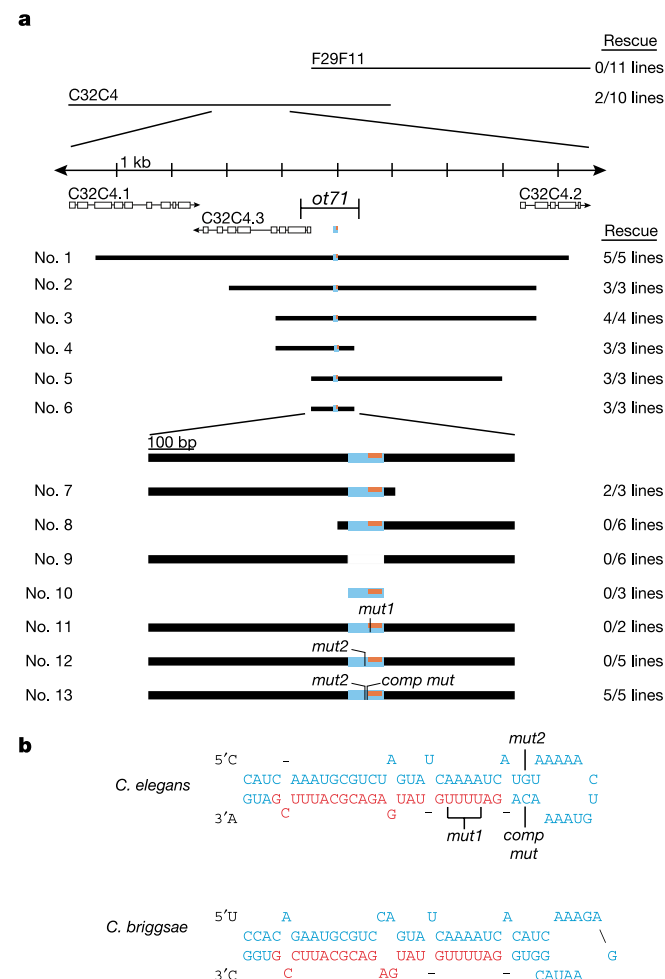


Figure 2 *Lsy-6* encodes a novel miRNA. **a**, Transformation rescue of the *lsy-6* mutant phenotype monitored by *gcy-5^{prom}::gfp* expression. (See Supplementary Information for rescue assay.) The predicted miRNA hairpin is shown in blue, the 23-nucleotide miRNA in red. *mut1* is a UUUU to GGGG change, *mut2* is a G to C change, and *comp mut* is the compensatory C to G change in the opposite strand of the hairpin. **b**, Structure of the conserved *lsy-6* hairpin in *C. elegans* and *C. briggsae*. The *lsy-6* miRNA is highlighted in red.

region of the *lsy-6* hairpin (*lsy-6^{hairpin}*), which contain the regulatory elements required for rescue of the *lsy-6* mutant phenotype, were fused to *gfp*. Although reporter gene fusions can only be an approximation of endogenous gene expression patterns, we were intrigued to observe that adult transgenic animals expressing this reporter show only limited expression in one pair of tail neurons and in about seven head neurons (Fig. 3a). Of the head neurons, expression was consistently observed in ASEL but not in ASER in adult animals (Fig. 3b). To confirm the cell specificity of *lsy-6* action and the sufficiency to promote ASEL cell fate, we expressed *lsy-6* under the control of the *ceh-36* promoter (*ceh-36^{prom}*), which is expressed bilaterally in post-mitotic ASEL and ASER^{3,19}. Bilateral expression of *lsy-6* in ASEL and ASER in a *lsy-6* null mutant background is sufficient to not only restore *gcy-7* expression and repress ectopic *gcy-5* gene expression in ASEL, but also to repress normal *gcy-5* expression and induce *gcy-7* expression in ASER, thus illustrating the instructive nature of *lsy-6* function in determining asymmetrical chemoreceptor expression profiles (Fig. 3c, d). Reversing the orientation of the hairpin abolishes ASEL fate-inducing activity (Fig. 3c).

We have previously identified an antagonistic relationship between transcriptional repressor and activator proteins as part of a regulatory cascade that determines left/right asymmetry in ASE cells (Fig. 4g)^{3,9}. *cog-1*, an *Nkx6*-type homeobox transcription factor,

acts with the *groucho*-like transcriptional co-repressor *unc-37* to repress ASEL cell fate in ASER. This is achieved, at least in part, through the repression of the *lim-6* homeobox gene, which represses ASER fate (Fig. 4g)³. *ceh-36*, an OTX-type homeodomain transcription factor, and *lin-49*, a transcriptional co-activator, act to promote the left cell fate in ASEL (Fig. 4g)³. To order *lsy-6* into this regulatory hierarchy, we first analysed the expression of the three known cell-specific transcription factors *ceh-36*, *lim-6*, and *cog-1* in a *lsy-6*(*ot71*) null mutant background. Expression of the bilaterally expressed *ceh-36* gene was unaffected, confirming that the overall features of ASE cell fate are unaffected (Fig. 4b, e). Consistent with the conversion of ASEL to ASER cell fate, expression of the normally left-expressed *lim-6* gene is lost (Fig. 4a). Conversely, *cog-1*, a negative regulator of *lim-6*, which is normally expressed more strongly in ASER than in ASEL³, is derepressed in ASEL (Fig. 4b).

As *cog-1* is a negative regulator of *lim-6* in ASER³ and as its expression is derepressed in ASEL in *lsy-6* mutants, loss of *lim-6* expression in *lsy-6* mutants might be due to ectopic *cog-1* expression in ASEL. To test this notion experimentally, we removed *cog-1* activity in a *lsy-6* null mutant background and observed derepression of *lim-6* expression (Fig. 4c). We conclude that *lsy-6* acts through *cog-1* to affect *lim-6* expression (Fig. 4g).

We next asked whether *cog-1* is a direct target of *lsy-6*. Examining the 3' UTR of *cog-1*, we noted a single site with significant complementarity to *lsy-6* (Fig. 4d; a second site with significantly lower similarity is also present; data not shown). In spite of an overall lack of sequence similarity of the whole 3' UTR, this putative *lsy-6* target site shows marked conservation to a site in the 3' UTR of the *cog-1* orthologue in the related nematode *C. briggsae* (Fig. 4d), implying functional relevance of this site. Moreover, the site has strongest complementarity to the 5' region of the miRNA, recently described as a common feature of miRNA interactions with their cognate targets²⁰.

To test whether the *cog-1* 3' UTR is indeed targeted by *lsy-6*, we generated *gfp* reporters with heterologous 3' UTRs. As shown previously, a *ceh-36^{prom}::gfp::unc-54^{3'UTR}* reporter construct is expressed bilaterally in ASEL and ASER³. We substituted the *unc-54* 3' UTR with the *cog-1* 3' UTR and examined whether endogenous *lsy-6* was sufficient to repress GFP expression in ASEL by means of the *cog-1* 3' UTR. In nematodes carrying the *ceh-36^{prom}::gfp::cog-1^{3'UTR}* transgene, reporter expression in ASEL is downregulated compared with ASER in the same animal and when compared with ASEL in control animals (Fig. 4e). Downregulation of the reporter in ASEL is dependent on the presence of *lsy-6*, as reporter gene expression in ASEL of *lsy-6* mutants is derepressed to levels that are indistinguishable from the levels in ASER (Fig. 4e). Moreover, downregulation of the reporter is dependent on the presence of the *lsy-6* complementary site, as deletion of this site from the reporter results in defective downregulation in ASEL (Fig. 4e). These results demonstrate that *lsy-6* acts through the *lsy-6* complementary site in the *cog-1* 3' UTR to repress gene expression.

To support further the notion of *lsy-6*-mediated regulation of *cog-1* expression, we asked whether *lsy-6* alone is sufficient to repress *cog-1* expression in cells that normally express *cog-1* but not *lsy-6*. We ectopically expressed the *lsy-6* hairpin under the control of the *cog-1* promoter (*cog-1^{prom}::lsy-6^{hairpin}*)—which is normally active in several head neurons, uterine cells and vulval cells²¹—and found that ectopic *lsy-6* expression can significantly reduce if not completely abolish *cog-1* expression (Fig. 4f). Expression of a mutated version of *lsy-6* does not show comparable downregulation of *cog-1* expression (Fig. 4f). Taken together, these findings indicate that *lsy-6* is both necessary and sufficient for the regulation of *cog-1* expression.

We ascribe the failure to identify the *lsy-6* miRNA in previously reported genome-wide miRNA searches^{14–17} to the following observations. First, although the *C. elegans* and *C. briggsae* *lsy-6* loci are

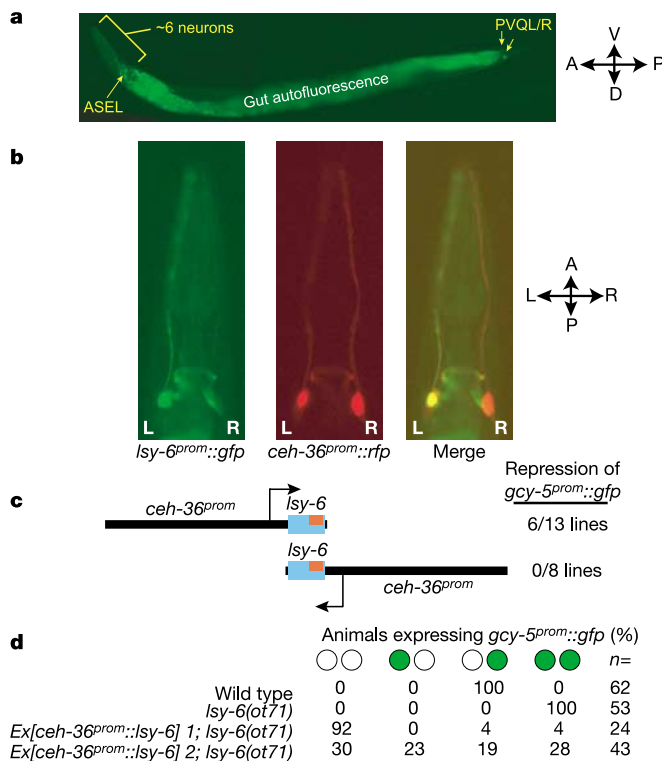


Figure 3 *Lsy-6* acts in a highly tissue-specific manner. **a**, Adult transgenic worm expressing *lsy-6^{prom}::gfp*. Neuronal cells expressing *lsy-6^{prom}::gfp* are: ASEL (see **b**), six radially symmetrical head sensory neurons and the PVQL/R tail interneurons. Autofluorescence of gut granules and posterior intestinal background fluorescence are signals in the mid-body region that are not reflective of *lsy-6* expression. Six of six transgenic lines show the same expression pattern. **b**, *Lsy-6* is expressed in ASEL but not ASER in adult animals. ASEL and ASER are labelled with a *ceh-36^{prom}::rfp* reporter construct (*otls151*). **c**, The *lsy-6* hairpin driven bilaterally in ASEL/R by the *ceh-36* promoter (*ceh-36^{prom}*) rescues the *lsy-6*(*ot71*) mutant phenotype in ASEL, as assayed by *gcy-5^{prom}::gfp* expression. **d**, Quantification of the effects of *ceh-36^{prom}::lsy-6^{hairpin}* expression on ASEL and ASER fate in *lsy-6* mutant animals. Two representative lines are shown. A, anterior; D, dorsal; P, posterior; V, ventral.

each predicted to form hairpin structures, their degree of overall conservation, particularly outside the conserved 23-nucleotide predicted miRNA, is limited, violating previous assertions on miRNA properties. Second, the potentially very restricted expression of *lsy-6* has probably hampered previous exhaustive efforts to clone all miRNAs by biochemical methods from total worm RNA. Our findings thus question previous claims on the completeness of miRNA complements within genomes¹⁵, and rather

suggest a previously unanticipated depth of miRNA abundance and function.

lsy-6 is the first miRNA identified to have a role in nervous system development *in vivo* and, moreover, provides novel insights into left/right patterning (Fig. 4g). We have shown previously that the ASE neurons are extremely sensitive to the correct dosage of *cog-1*, such that extra copies of *cog-1* readily produce left/right asymmetry defects³. Transcriptional mechanisms may not be sufficient to

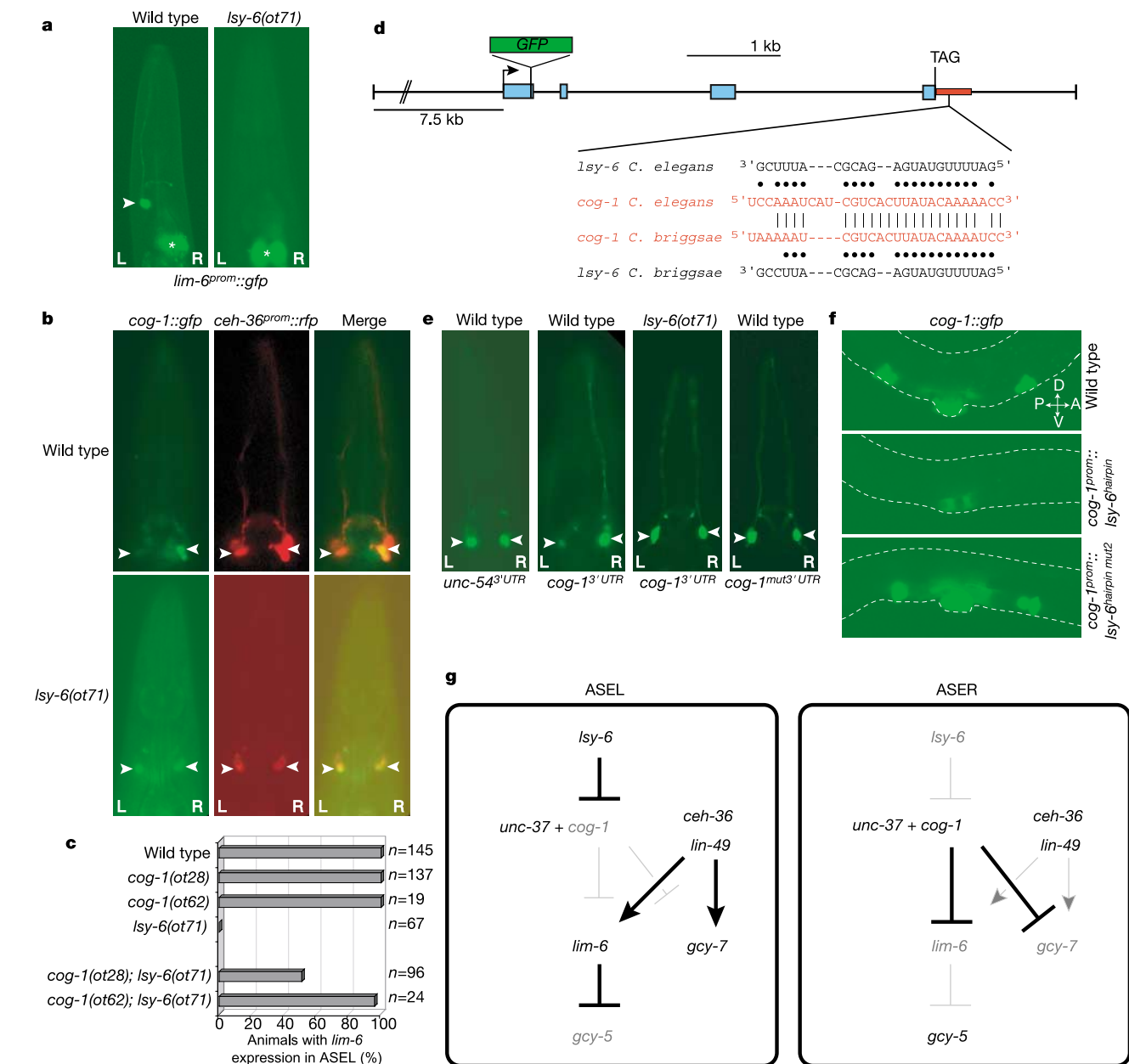


Figure 4 *Lsy-6* represses *cog-1* to promote ASEL cell fate. **a**, Loss of *lsy-6* causes loss of expression of the *lim-6* homeobox gene in ASEL (white arrowhead; asterisks indicate expression in the excretory gland cells). (See Supplementary Information for quantification of effects.) **b**, Loss of *lsy-6* causes loss of *cog-1* repression in ASEL and leaves *ceh-36* expression in ASEL/R unaffected (white arrowheads; *ceh-36* is also expressed in AWCL/R¹⁹). *cog-1* expression is monitored with a translational *gfp* reporter array, *sys63* (ref. 21), which is schematically shown in **d**. (See Supplementary Information for quantification of effects.) **c**, *Lsy-6* acts through *cog-1* to affect *lim-6* expression. *ot28* is a weak hypomorphic allele; *ot62* is a strong loss of function allele³. **d**, The *cog-1* 3' UTR contains a *lsy-6* complementary site. The *cog-1::gfp* reporter construct used in **b** and **f** is

shown, as well as its 3' UTR (red box)²¹ and the *lsy-6* complementary site in the 3' UTR. A second, significantly weaker *lsy-6* complementary site can be found at the extreme end of the 3' UTR (not shown). Black dots indicate complementarity. **e**, *Lsy-6* acts through the *lsy-6* complementary site in the *cog-1* 3' UTR. Heterologous 3' UTRs were fused 3' to the *gfp* coding sequence driven by the *ceh-36* promoter, and expression was analysed in wild type and *lsy-6* mutants. (See Supplementary Information for quantification of effects.) **f**, Repression of *cog-1* expression (monitored with the *sys63* transgene; construct schematically shown in **d**) through ectopic *lsy-6* expression in the uterus and vulva. (See Supplementary Information for quantification of effects.) **g**, Genetic pathway leading to asymmetrical *gcy* gene expression in ASEL and ASER⁸.

regulate precise levels of *cog-1* gene expression, thus necessitating an additional level of gene expression control mediated by a miRNA. The elucidation of mechanisms that restrict *lxy-6* expression to just one of two bilaterally symmetrical taste neurons will provide further insights into the molecular mechanisms of establishing left/right asymmetry. □

Methods

Wild-type and mutant strains

We used the following nematode strains: wild-type N2 variation Bristol, CB4856 Hawaiian wild-type isolate, OH2535: *lxy-6*(*ot71*), OH153: *cog-1*(*ot28*), and OH1445: *cog-1*(*ot62*)/+; *ot1s114*; *him-5*(*e1490*).

Reporter transgenes

We used the following reporter transgenes: *ntIs1* *Is*[*gcy-5^{prom}::gfp*; *lin-15*(+)]³, *otIs3* *Is*[*gcy-7^{prom}::gfp*; *lin-15*(+)]³, *otIs114* *Is*[*lim-6^{prom}::gfp*; *rol-6*(*d*)]³, *otIs151* *Is*[*ceh-36^{prom}::rfp*; *rol-6*(*d*)], and *syls63* *Is*[*cog-1::gfp*; *dpy-20*(+)]²¹. (See Supplementary Information for a description of other transgenic lines.)

DNA construction and injection

All constructs were generated by polymerase chain reaction fusion²². A list of all constructs and primers used can be found in the Supplementary Information. DNA was injected at 2–20 ng μl^{-1} depending on the experiment (see Supplementary Information) with either *rol-6* (100 ng μl^{-1}) or *unc-122::gfp* (50 ng μl^{-1}) as the injection marker.

Scoring of phenotype

Animals were scored as adults. Quantification of defects shown in the figures can be found in the Supplementary Information.

Received 2 November; accepted 1 December 2003; doi:10.1038/nature02255.

Published online 14 December 2003.

- Yu, S., Avery, L., Baude, E. & Garbers, D. L. Guanylyl cyclase expression in specific sensory neurons: a new family of chemosensory receptors. *Proc. Natl Acad. Sci. USA* **94**, 3384–3387 (1997).
- Pierce-Shimomura, J. T., Fautmont, S., Gaston, M. R., Pearson, B. J. & Lockery, S. R. The homeobox gene *lim-6* is required for distinct chemosensory representations in *C. elegans*. *Nature* **410**, 694–698 (2001).
- Chang, S., Johnston, R. J. & Hobert, O. A transcriptional regulatory cascade that controls left/right asymmetry in chemosensory neurons of *C. elegans*. *Genes Dev.* **17**, 2123–2137 (2003).
- Halpern, M. E., Liang, J. O. & Gamse, J. T. Leaning to the left: laterality in the zebrafish forebrain. *Trends Neurosci.* **26**, 308–313 (2003).
- Yost, H. J. Left-right development from embryos to brains. *Dev. Genet.* **23**, 159–163 (1998).
- Hobert, O., Johnston, R. J. & Chang, S. Left-right asymmetry in the nervous system: the *Caenorhabditis elegans* model. *Nature Rev. Neurosci.* **3**, 629–640 (2002).
- Bargmann, C. I. & Horvitz, H. R. Chemosensory neurons with overlapping functions direct chemotaxis to multiple chemicals in *C. elegans*. *Neuron* **7**, 729–742 (1991).
- White, J. G., Southgate, E., Thomson, J. N. & Brenner, S. The structure of the nervous system of the nematode *Caenorhabditis elegans*. *Phil. Trans. R. Soc. Lond. B* **314**, 1–340 (1986).
- Hobert, O., Tessmar, K. & Ruvkun, G. The *Caenorhabditis elegans* *lim-6* LIM homeobox gene regulates neurite outgrowth and function of particular GABAergic neurons. *Development* **126**, 1547–1562 (1999).
- Troemel, E. R., Sagasti, A. & Bargmann, C. I. Lateral signaling mediated by axon contact and calcium entry regulates asymmetric odorant receptor expression in *C. elegans*. *Cell* **99**, 387–398 (1999).
- Wicks, S. R., Yeh, R. T., Gish, W. R., Waterston, R. H. & Plasterk, R. H. Rapid gene mapping in *Caenorhabditis elegans* using a high density polymorphism map. *Nature Genet.* **28**, 160–164 (2001).
- Lee, R. C., Feinbaum, R. L. & Ambros, V. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell* **75**, 843–854 (1993).
- Olsen, P. H. & Ambros, V. The *lin-4* regulatory RNA controls developmental timing in *Caenorhabditis elegans* by blocking LIN-14 protein synthesis after the initiation of translation. *Dev. Biol.* **216**, 671–680 (1999).
- Ambros, V., Lee, R. C., Lavanway, A., Williams, P. T. & Jewell, D. MicroRNAs and other tiny endogenous RNAs in *C. elegans*. *Curr. Biol.* **13**, 807–818 (2003).
- Lim, L. P. et al. The microRNAs of *Caenorhabditis elegans*. *Genes Dev.* **17**, 991–1008 (2003).
- Grad, Y. et al. Computational and experimental identification of *C. elegans* microRNAs. *Mol. Cell* **11**, 1253–1263 (2003).
- Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294**, 858–862 (2001).
- Johnson, S. M., Lin, S. Y. & Slack, F. J. The time of appearance of the *C. elegans* *let-7* microRNA is transcriptionally controlled utilizing a temporal regulatory element in its promoter. *Dev. Biol.* **259**, 364–379 (2003).
- Lanjuin, A., VanHoven, M. K., Bargmann, C. I., Thompson, J. K. & Sengupta, P. Otx/otd homeobox genes specify distinct sensory neuron identities in *C. elegans*. *Dev. Cell* **5**, 621–633 (2003).
- Stark, A., Brennecke, J., Russell, R. B. & Cohen, S. M. Identification of *Drosophila* MicroRNA targets. *PLOS Biol.* Online publication 13 October 2003 (doi:10.1371/journal.pbio.0000060).
- Palmer, R. E., Inoue, T., Sherwood, D. R., Jiang, L. I. & Sternberg, P. W. *Caenorhabditis elegans* *cog-1* locus encodes GTX/Nkx6.1 homeodomain proteins and regulates multiple aspects of reproductive system development. *Dev. Biol.* **252**, 202–213 (2002).
- Hobert, O. PCR fusion-based approach to create reporter gene constructs for expression analysis in transgenic *C. elegans*. *Biotechniques* **32**, 728–730 (2002).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are grateful to A. Wenick and B. Reinhart for suggestions; Q. Chen for technical assistance; R. Mann, L. Johnston, P. Sengupta and A. Lanjuin for critically reading the manuscript; and the NSF (R.J.J.) and NIH (O.H.) for funding.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to O.H. (or38@columbia.edu).

A self-organizing system of repressor gradients establishes segmental complexity in *Drosophila*

Dorothy E. Clyde^{*†}, Maria S. G. Corado^{*}, Xuelin Wu[†], Adam Paré, Dmitri Papatsenko & Stephen Small

Biology Department, New York University, New York, New York 10003, USA

* These authors contributed equally to this work

† Present addresses: Biosciences Department, University of Kent, Canterbury CT2 7NJ, UK (D.E.C.); Plant Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, California 92037, USA (X.W.)

Gradients of regulatory factors are essential for establishing precise patterns of gene expression during development^{1–3}; however, it is not clear how patterning information in multiple gradients is integrated to generate complex body plans. Here we show that opposing gradients of two *Drosophila* transcriptional repressors, Hunchback (Hb) and Knirps (Kni), position several segments by differentially repressing two distinct regulatory regions (enhancers) of the pair-rule gene *even-skipped* (*eve*). Computational and *in vivo* analyses suggest that enhancer sensitivity to repression is controlled by the number and affinity of repressor-binding sites. Because the *kni* expression domain is positioned between two gradients of Hb, each enhancer directs expression of a pair of symmetrical stripes, one on each side of the *kni* domain. Thus, only two enhancers are required for the precise positioning of eight stripe borders (four stripes), or more than half of the whole *eve* pattern. Our results show that complex developmental expression patterns can be generated by simple repressor gradients. They also support the utility of computational analyses for defining and deciphering regulatory information contained in genomic DNA.

In *Drosophila*, the pair-rule gene *eve* is expressed in a pattern of seven stripes during the syncytial blastoderm stage of development. This pattern foreshadows the mature segmented body plan and is regulated by five enhancers^{4–8}. Three enhancers drive expression of single stripes (*eve* 1, *eve* 2 and *eve* 5), and the remaining two drive expression of pairs of stripes (*eve* 3 + 7 and *eve* 4 + 6). The best characterized *eve* enhancer drives the expression of stripe 2 (*eve* 2)^{9,10}, which is activated in a broad anterior domain by the maternal morphogens Bicoid and Hb. Borders of the stripe are formed by repressive interactions involving the gap proteins Giant (Gt) and Kruppel (Kr), which are expressed in gradients anterior and posterior to the stripe, respectively. Activation and repression are mediated by the direct binding of all four proteins to discrete sites in the enhancer^{9,11}. Thus, this enhancer acts as a transcriptional switch that senses activator/repressor ratios in individual nuclei.

Considerably less is known about the molecular regulation of the enhancers that drive two stripes. *eve* 3 + 7 is activated by ubiquitous factors including dSTAT92E^{12,13}, and activation of *eve* 4 + 6 requires the function of the *fish-hook* gene¹⁴, but other activators are unknown. Genetic studies showed that the gap genes *hb* and *kni* are required for forming the borders of all four of these stripes. *kni* is

expressed in a broad posterior domain located between *eve* stripes 4 and 6 (Fig. 1a). In *kni* mutants, the two-stripe patterns driven by *eve* 3 + 7-*lacZ* and *eve* 4 + 6-*lacZ* reporter genes are completely repressed in the region between the stripes^{6,8}. By contrast, *hb* is expressed in an anterior domain that abuts *eve* 3 and a broad posterior stripe that overlaps *eve* 7 (Fig. 1b). In zygotic *hb* mutants, there are marked derepressions of the outer borders of the stripes driven by both the *eve* 3 + 7 and *eve* 4 + 6 reporter genes^{6,8}.

To test whether the *eve* 3 + 7 and *eve* 4 + 6 enhancers are differentially sensitive to Kni- and Hb-mediated repression, we used the *snail*(*sna*) promoter to misexpress these genes along the

ventral surface of the embryo (Fig. 1f, k). The ectopic domain directed by this promoter is uniformly distributed along the anterior-posterior axis¹⁵, and forms a ventral to dorsal gradient of protein diffusion (data not shown). As all seven *eve* stripes are subject to the same increase in protein concentration, differential sensitivities among stripes can be assayed directly. Weakly affected stripes will be repressed only in the ventral-most nuclei, whereas strongly affected stripes will show repression in more lateral or even dorsal regions.

Ventral expression of either Kni (*sna:kni*) or Hb (*sna:hb*) is sufficient for repression of *eve* stripes 3, 4, 6 and 7 in ventral regions

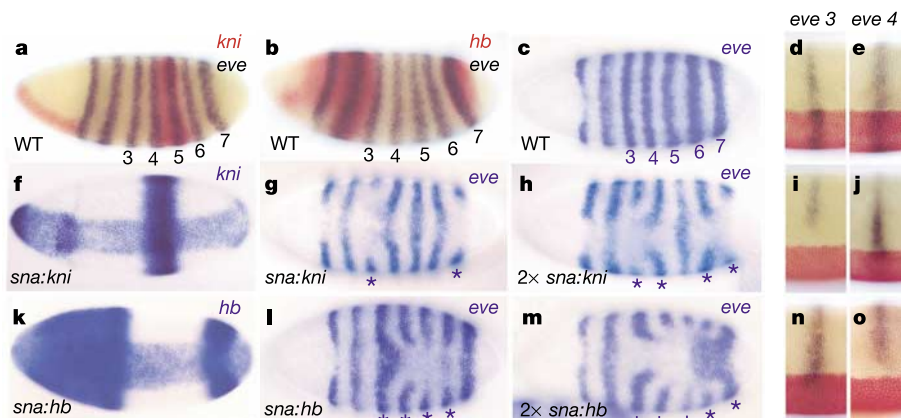


Figure 1 Individual *eve* stripes are differentially responsive to gradients of Kni and Hb. **a-e**, mRNA expression patterns in wild-type embryos at mid-cleavage cycle 14. **a, b**, Lateral views showing the spatial relationships between *hb*, *kni* and *eve*. **c**, Ventral view of the expression pattern of *eve*. Stripes 3-7 are numbered. **d, e**, Lateral views of embryos carrying the *eve* 3 + 7-*lacZ* (**d**) or the *eve* 4 + 6-*lacZ* (**e**) transgenes. Only the regions of stripe 3 or stripe 4 (black) are shown. Embryos are also stained to detect the

expression pattern of endogenous *sna* (red). **f-o**, Expression patterns in embryos containing an ectopic domain of Kni (**f-j**) or Hb (**k-o**) expression along the ventral embryonic surface. **f-h, k-m**, Ventral views showing stripe-specific repression (compare with **c**). Affected stripes are marked by asterisks. **i, j, n, o**, Lateral views of embryos carrying the *eve* 3 + 7-*lacZ* (**i, n**) or the *eve* 4 + 6-*lacZ* transgene (**j, o**), in addition to the *sna:kni* or *sna:hb* misexpression constructs (compare with **d** and **e**).

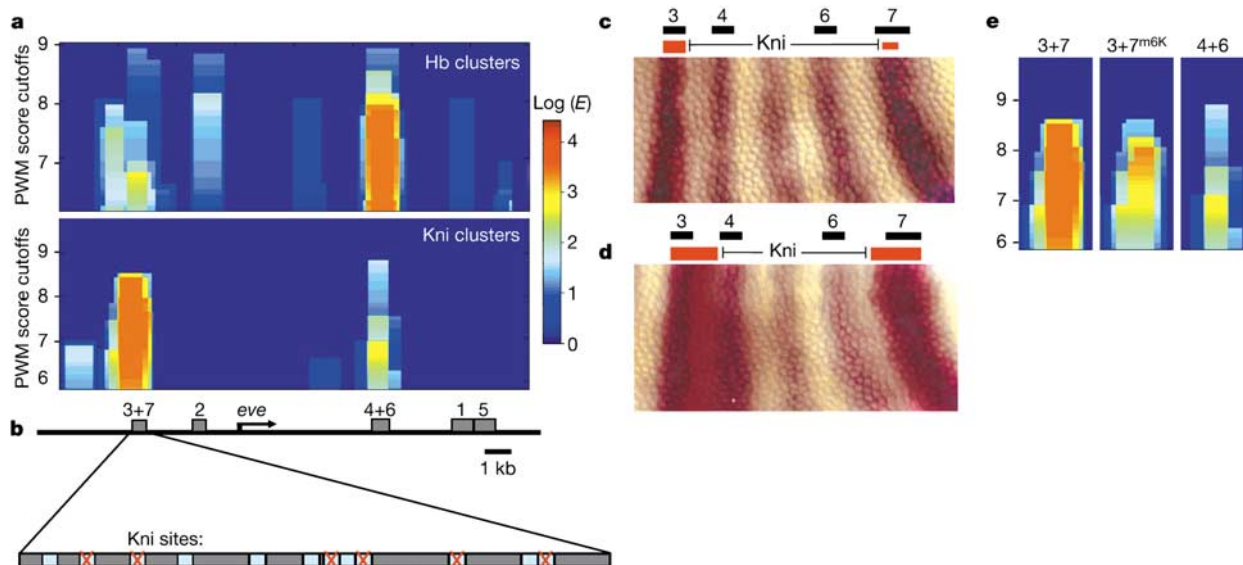


Figure 2 Clusters of repressor-binding sites determine enhancer sensitivity. **a**, Computer-predicted Hb- and Kni-binding site clusters in the *eve* locus. The heights of the bars along the y axis represent the binding affinities (PWM values) of the clustered sites. Site density is represented by a colour code (right). The positions of the clusters in the *eve* locus are represented along the x axis. **b**, The 500-bp minimal *eve* 3 + 7 enhancer contains 12 predicted Kni-binding sites. Six of these sites (X) were mutated to reduce their PWM value to close to zero (3 + 7m6K; see Supplementary Table 1).

c, d, Expression of endogenous *eve* mRNA (black) compared with *lacZ* reporter mRNA (red) driven by the wild-type 3 + 7 enhancer (**c**) and the mutant 3 + 7m6K enhancer (**d**). Note the derepression of the 3 + 7m6K-*lacZ* reporter as compared with 3 + 7-*lacZ*. **e**, Mutating 6 of the 12 Kni sites in the 3 + 7 enhancer changes the apparent 'sensitivity' of the enhancer to one that is intermediate between the wild-type 3 + 7 and 4 + 6 enhancers.

(Fig. 1), but specific stripes require different quantities of ectopic protein for repression. One copy of the *sna:kni* transgene ($1 \times \text{sna:kni}$) represses *eve* stripes 3 and 7, but has little effect on stripes 4 and 6 (Fig. 1g). Two copies repress all four stripes, but stripes 3 and 7 are more strongly repressed than stripes 4 and 6 (Fig. 1h). Misexpression of Hb shows the opposite effects. One copy of *sna:hb* causes a strong repression of stripes 4, 5 and 6, and an anterior weakening and posterior expansion of stripe 3 (Fig. 1l). The posterior expansion is probably caused by Hb-mediated repression of *kni* (see below).

Two copies of *sna:hb* cause a stronger repression of stripes 4, 5 and 6, repress stripe 3 completely in ventral-most nuclei, and considerably affect stripe 7, which seems slightly weaker and expanded anteriorly, again toward the region normally occupied by *kni* (Fig. 1m). The weaker effect on stripe 7 suggests that higher concentrations of Hb are required to repress this stripe. This is consistent with the fact that the posterior *hb* stripe overlaps stripe 7, and that additional factors (including Tll) are required for activation of this stripe⁸. The strong repressive effect of ectopic Hb on stripe 5 (Fig. 1l, m) is unexpected as this stripe seems to be normal in *hb* mutants⁶. In addition, computational analysis shows that there are very few Hb-binding sites in the *eve* 5 enhancer region (Fig. 2a). These results suggest that Hb-mediated repression of this stripe is indirect.

The above results suggest that the *eve* 3 + 7 and *eve* 4 + 6 enhancers respond autonomously to different amounts of the Hb and Kni repressors. To test this idea further, *lacZ* reporter genes driven by the minimal *eve* 3 + 7 or *eve* 4 + 6 enhancer were crossed into embryos carrying the *sna:kni* or *sna:hb* misexpression transgene (Fig. 1i, j, n, o). Embryos were also stained for endogenous *sna* expression, which forms a sharp ventral–lateral border, a landmark for measuring the extent of repression along the dorsal–ventral axis.

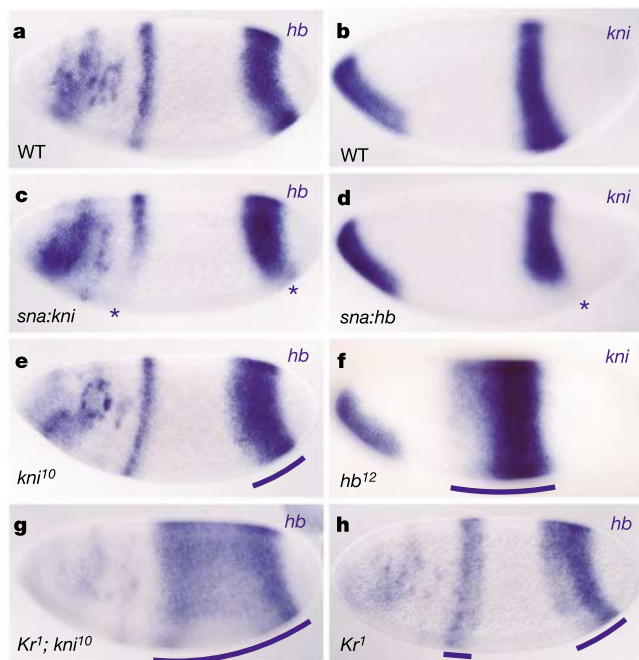


Figure 3 Mutual repression between Hb and Kni. **a**, *hb* mRNA expression in a wild-type embryo. **b**, *kni* mRNA expression in a wild-type embryo. **c**, *hb* mRNA expression in an embryo carrying one copy of the *sna:kni* transgene. **d**, *kni* mRNA expression in an embryo carrying one copy of the *sna:hb* transgene. **e**, *hb* mRNA expression in a *kni*¹⁰ mutant embryo. **f**, *kni* mRNA expression in a *hb*¹² mutant embryo. **g**, *hb* mRNA expression in a *Kr*¹ *kni*¹⁰ double mutant embryo. **h**, *hb* mRNA expression in a *Kr*¹ mutant embryo. Repressive activities are indicated by asterisks, domain expansions are marked by lines below the embryos.

Ventral repression of the *eve* 3 + 7-*lacZ* transgene by Kni ($2 \times \text{sna:kni}$) extends at least five nuclei above the *sna* border, but the *eve* 4 + 6-*lacZ* transgene is repressed only within the *sna* domain (Fig. 1i, j). Ventral expression of Hb ($1 \times \text{sna:hb}$) causes the opposite effects: the *eve* 4 + 6-*lacZ* transgene is more strongly repressed than *eve* 3 + 7-*lacZ* (Fig. 1n, o). These experiments are consistent with the effects observed for the endogenous *eve* stripes.

To determine how these enhancers sense differences in repressor concentration, we used bioinformatics to analyse the distribution and affinity of Hb- and Kni-binding sites in the *eve* locus. Position-weighted matrices (PWMs) for each protein were generated by compiling and aligning the sequences of all known Hb- and Kni-binding sites¹⁶, and a clustering algorithm¹⁷ was used to search the 20-kilobase (kb) region surrounding the *eve* locus. This analysis identified only two main clusters for each factor in this region, which overlap precisely with the positions of the *eve* 3 + 7 and 4 + 6 enhancers (Fig. 2a). The composition of sites within these clusters, however, is very different. The 3 + 7 enhancer contains considerably more Kni sites with higher PWM scores than does the 4 + 6 enhancer (Supplementary Figs 1 and 2), consistent with its higher sensitivity to repression by Kni.

For Hb, searching with a low PWM cutoff value (>4.0) identified 11 sites in the more sensitive 4 + 6 enhancer and, unexpectedly, 16 sites in the 3 + 7 enhancer. These results are similar to previous findings¹⁸; however, 10 of the 11 Hb sites in the 4 + 6 enhancer have very high PWM scores (>6.3), as compared with only 6 in the 3 + 7 enhancer. Also, six of the ten high-scoring sites in the 4 + 6 enhancer are very tightly clustered in a 130-base-pair (bp) interval, whereas those in the 3 + 7 enhancer are evenly distributed across the sequence. These results suggest that binding-site affinity and distribution may be crucial parameters in determining enhancer sensitivity to Hb-mediated repression.

We next tested whether the clustering algorithm could predictably change enhancer sensitivity, using the Kni-binding sites in the 3 + 7 enhancer as a test case. The PWM search identified 12 Kni-binding sites in the minimal 3 + 7 enhancer; six of these sites were mutated (Supplementary Table 1) so that the cluster significance score of the mutated enhancer (denoted 3 + 7^{m6K}) was intermediate between those of the wild-type 3 + 7 and 4 + 6 enhancers (Fig. 2b, e). Reporter expression driven by 3 + 7^{m6K} shows a derepression of the inner borders of stripes 3 and 7, suggesting that Kni-mediated repression has been compromised by these mutations. The stripe 3 response of the mutated enhancer extends throughout the inter-stripe region posterior to *eve* 3 to the anterior border of, but not through the region occupied by, *eve* 4 (Fig. 2, compare c and d). Thus, the 3 + 7^{m6K} enhancer is less sensitive to Kni than is the wild-

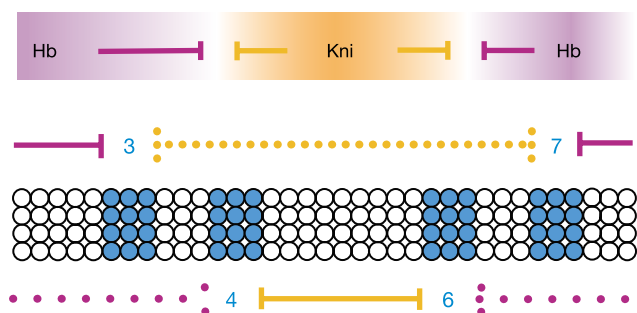


Figure 4 Repressor gradients and the generation of pattern complexity. Mutual repression between the Hb and Kni gradients fixes the positions of, and spacing between, their repressor gradients. These gradients are then 'read' by two enhancers that are differentially sensitive to repression by Hb or Kni, which ensures that *eve* stripes 3, 4, 6 and 7 are expressed at precisely defined positions.

type 3 + 7 enhancer, but is still more sensitive than the 4 + 6 enhancer. This suggests that the precise positioning of these stripes is controlled by the 'strength' of Kni site clusters.

As the normal Hb and Kni gradients set several expression boundaries in the region between their domains, it is essential that their relative positions in the embryo are precisely established and maintained. This could be achieved by mutual repression, as suggested by previous studies^{2,19,20}. To test this further, we analysed the effects of ventrally expressed Kni on the expression of *hb* messenger RNA, and vice versa. Misexpression of Kni causes a strong reduction in *hb* mRNA in ventral regions (compare Fig. 3a and Fig. 3c). Similarly, misexpressed Hb causes a strong repression of *kni* (compare Fig. 3b and Fig. 3d).

Loss-of-function experiments lend further weight to the mutual repression hypothesis. In *hb* mutants, there is a substantial expansion of the posterior *kni* domain (Fig. 3f). In *kni* mutants, there is a slight anterior expansion of the posterior *hb* domain, but no effect on the anterior domain. Double mutant embryos that lack *kni* and the central gap gene *Kruppel* (*Kr*) show, however, a marked expansion of zygotic *hb* expression throughout the posterior half of the embryo (Fig. 3e, g, h). Because misexpression of *Kr* alone has no effect on the *hb* expression pattern (data not shown), this observation suggests that *Kr* and Kni may cooperate in repression of *hb*.

In conclusion, we have demonstrated the principle elements of a simple repression system that greatly increases pattern complexity in the *Drosophila* embryo. Strong reciprocal repression between *kni* and *hb* positions a symmetrical Kni domain between two opposing gradients of Hb (Fig. 4). This arrangement permits a single enhancer to make two stripes, one on both sides of the Kni domain. Two differentially sensitive enhancers effectively double the patterning information in each gradient, leading to the establishment of eight expression boundaries. A similar antagonistic relationship exists between the gap genes *gt* and *Kr*, which are expressed in non-overlapping domains, with the central *Kr* domain positioned between two *gt* domains²¹. The *eve* 2 and *eve* 5 stripes are formed on either side of the *Kr* domain by *Kr*- and *Gt*-mediated repression, but in this case each stripe is regulated by a separate enhancer, probably because the activators of these stripes are expressed in localized patterns^{6,7}.

Previous studies have shown that activator gradients are crucial for differential positioning of target gene expression patterns along the anterior–posterior and dorsal–ventral axes^{3,22,23}. Our results and other studies^{2,24} suggest that repressor gradients can also specify several gene expression boundaries by interacting with differentially sensitive regulatory elements. At the molecular level, repression mechanisms are flexible: enhancer activation can be prevented by direct repression or by interfering with the binding or activity of even a single activator protein^{25,26}. We propose that repressor gradients, owing to this flexibility, are inherently more effective than activator gradients at providing developmental patterning information.

The sensitivity of an enhancer is likely to be determined by several parameters including the number, affinity and arrangement of repressor-binding sites, but predicting the relative importance of each of these parameters for a given enhancer is difficult. For the Kni repressor gradient, the different responses of the 3 + 7 and 4 + 6 enhancers seem to depend on different numbers of binding sites. By contrast, the different responses of the same enhancers to Hb repression seem to depend on the affinity and/or arrangement of sites. Thus, it may be impossible to formulate simple rules that describe the functional characteristics of most enhancers. However, future studies that combine computational analyses with experimental tests will undoubtedly increase our ability to identify and to characterize the genomic elements that regulate transcription. □

Methods

Fly stocks

We used the following amorphic alleles: *kni*¹⁰, *hb*¹² and *Kr*¹. The *eve* 4 + 6–*lacZ* line (PC N40C52-D⁶) was a gift from M. Fujioka and J. Jaynes. The *eve* 3 + 7–*lacZ* line used here contains the 500-bp minimal *eve* 3 + 7 enhancer and has been described⁸.

Ventral expression of Kni and Hb

To make the *sna:kni* transgene, the *knirps* coding sequence and *eve* 3' untranslated region (UTR) were excised as an *Asp*718–*Xba*I fragment from construct 22FKE¹⁹, blunt-ended and cloned into the *Pme*I site of *pCas:sna* (ref. 27), a gift from L. Andrioli. For *sna:hb*, a 2.4-kilobase (kb) *Nde*I–*Eco*RI *hb* complementary DNA fragment was fused to an 0.8-kb *Eco*RI–*Xba*I fragment from the α -*tubulin* 3' UTR, blunt-ended and cloned into the *Pme*I site of *pCas:sna*. We generated 7 and 15 independent lines with the *sna:hb* and *sna:kni* constructs, respectively, as described²⁸. All misexpression lines tested generated quantities of ectopic protein that were considerably lower than endogenous Hb and Kni.

The *sna:hb* and *sna:kni* constructs contain transcriptional stop signals flanked by FRT sites downstream of the *sna* promoter to prevent ectopic expression during production and maintenance of transgenic lines. To activate ventral expression, males carrying a β 2-*tubulin*–FLP transgene²⁹ and *sna:hb* or *sna:kni* were crossed to *yw* virgins, or those carrying the *eve* 3 + 7–*lacZ* or *eve* 4 + 6–*lacZ* transgenes. Embryos aged 2–4 h were collected and analysed by *in situ* hybridization as described²⁸.

Mutagenesis of the *eve* 3 + 7–*lacZ* reporter construct

Kni-binding sites in the *eve* 3 + 7 enhancer were mutated by site-directed mutagenesis of single-stranded DNA using a pBS-3 + 7NKS construct as a template. We used the following oligomers: N2, 5'–CAAAAACTGATCTAGCTAGCTAATGCAGCGAG–3'; N3, 5'–GCTGGGAAATGGCTAGGCGGCCATAAACCG–3'; N8 + 9, 5'–GCGCACAAATGGCTAGAAAAACGTGATCTACCTAGCTAATACGGGCG–3'; N10, 5'–CGCTGGGTTTGGGCTAGAAAACTAGCGCAG–3'; N11, 5'–CAACACAAACAAACGCTTGTAAAAACGAGAGC–3'.

The mutant enhancer (3 + 7^{m6K}) was cloned into *pCaSpeR–eve–lacZ*, which contains the *eve* basal promoter, the *lacZ* cDNA and the α -*tubulin* 3' UTR³⁰. The sequence of the minimal *eve* 3 + 7 enhancer showing the positions of all predicted Kni and Hb sites is provided as Supplementary Fig. 1. Four lines containing the 3 + 7^{m6K}–*lacZ* construct were generated and tested. All gave similar results.

Received 30 June; accepted 23 October 2003; doi:10.1038/nature02189.

1. Driever, W., Thoma, G. & Nusslein-Volhard, C. Determination of spatial domains of zygotic gene expression in the *Drosophila* embryo by the affinity of binding sites for the bicoid morphogen. *Nature* **340**, 363–367 (1989).
2. Struhl, G., Johnston, P. & Lawrence, P. A. Control of *Drosophila* body pattern by the hunchback morphogen gradient. *Cell* **69**, 237–249 (1992).
3. Jiang, J. & Levine, M. Binding affinities and cooperative interactions with bHLH activators delimit threshold responses to the dorsal gradient morphogen. *Cell* **72**, 741–752 (1993).
4. Goto, T., Macdonald, P. & Maniatis, T. Early and late periodic patterns of even-skipped expression are controlled by distinct regulatory elements that respond to different spatial cues. *Cell* **57**, 413–422 (1989).
5. Harding, K., Hoey, T., Warrior, R. & Levine, M. Autoregulatory and gap gene response elements of the even-skipped promoter of *Drosophila*. *EMBO J.* **8**, 1205–1212 (1989).
6. Fujioka, M., Emi-Sarker, Y., Yusibova, G. L., Goto, T. & Jaynes, J. B. Analysis of an even-skipped rescue transgene reveals both composite and discrete neuronal and early blastoderm enhancers, and multi-stripe positioning by gap gene repressor gradients. *Development* **126**, 2527–2538 (1999).
7. Small, S., Blair, A. & Levine, M. Regulation of even-skipped stripe 2 in the *Drosophila* embryo. *EMBO J.* **11**, 4047–4057 (1992).
8. Small, S., Blair, A. & Levine, M. Regulation of two pair-rule stripes by a single enhancer in the *Drosophila* embryo. *Dev. Biol.* **175**, 314–324 (1996).
9. Small, S., Kraut, R., Hoey, T., Warrior, R. & Levine, M. Transcriptional regulation of a pair-rule stripe in *Drosophila*. *Genes Dev.* **5**, 827–839 (1991).
10. Stanojevic, D., Small, S. & Levine, M. Regulation of a segmentation stripe by overlapping activators and repressors in the *Drosophila* embryo. *Science* **254**, 1385–1387 (1991).
11. Stanojevic, D., Hoey, T. & Levine, M. Sequence-specific DNA-binding activities of the gap proteins encoded by hunchback and Kruppel in *Drosophila*. *Nature* **341**, 331–335 (1989).
12. Yan, R., Small, S., Desplan, C., Dearolf, C. R. & Darnell, J. E. Jr Identification of a Stat gene that functions in *Drosophila* development. *Cell* **84**, 421–430 (1996).
13. Hou, X. S., Melnick, M. B. & Perrimon, N. Marelle acts downstream of the *Drosophila* HOP/JAK kinase and encodes a protein similar to the mammalian STATs. *Cell* **84**, 411–419 (1996); erratum *ibid.* **Cell** **85**, 290 (1996).
14. Nambu, P. A. & Nambu, J. R. The *Drosophila* fish-hook gene encodes a HMG domain protein essential for segmentation and CNS development. *Development* **122**, 3467–3475 (1996).
15. Ip, Y. T., Park, R. E., Kosman, D., Yazdanbakhsh, K. & Levine, M. *dorsal–twist* interactions establish *snail* expression in the presumptive mesoderm of the *Drosophila* embryo. *Genes Dev.* **6**, 1518–1530 (1992).
16. Papatsenko, D. A. *et al.* Extraction of functional binding sites from unique regulatory regions: the *Drosophila* early developmental enhancers. *Genome Res.* **12**, 470–481 (2002).
17. Lifanov, A. P., Makeev, V. J., Nazina, A. G. & Papatsenko, D. A. Homotypic regulatory clusters in *Drosophila*. *Genome Res.* **13**, 579–588 (2003).
18. Berman, B. P. *et al.* Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl Acad. Sci. USA* **99**, 757–762 (2002).
19. Kosman, D. & Small, S. Concentration-dependent patterning by an ectopic expression domain of the *Drosophila* gap gene *knirps*. *Development* **124**, 1343–1354 (1997).
20. Hulskamp, M., Pfeifle, C. & Tautz, D. A morphogenetic gradient of hunchback protein organizes the expression of the gap genes *Kruppel* and *knirps* in the early *Drosophila* embryo. *Nature* **346**, 577–580 (1990).

21. Kraut, R. & Levine, M. Mutually repressive interactions between the gap genes giant and Kruppel define middle body regions of the *Drosophila* embryo. *Development* **111**, 611–621 (1991).
22. Struhl, G., Struhl, K. & Macdonald, P. M. The gradient morphogen bicoid is a concentration-dependent transcriptional activator. *Cell* **57**, 1259–1273 (1989).
23. Driever, W. & Nusslein-Volhard, C. The bicoid protein determines position in the *Drosophila* embryo in a concentration-dependent manner. *Cell* **54**, 95–104 (1988).
24. Langeland, J. A., Attai, S. F., Vorwerk, K. & Carroll, S. B. Positioning adjacent pair-rule stripes in the posterior *Drosophila* embryo. *Development* **120**, 2945–2955 (1994).
25. Hanna-Rose, W. & Hansen, U. Active repression mechanisms of eukaryotic transcription repressors. *Trends Genet.* **12**, 229–234 (1996).
26. Gray, S. & Levine, M. Transcriptional repression in development. *Curr. Opin. Cell Biol.* **8**, 358–364 (1996).
27. Andrioli, L. P., Vasisht, V., Theodosopoulou, E., Oberstein, A. & Small, S. Anterior repression of a *Drosophila* stripe enhancer requires three position-specific mechanisms. *Development* **129**, 4931–4940 (2002).
28. Small, S. *In vivo* analysis of *lacZ* fusion genes in transgenic *Drosophila melanogaster*. *Methods Enzymol.* **326**, 146–159 (2000).
29. Struhl, G., Fitzgerald, K. & Greenwald, I. Intrinsic activity of the Lin-12 and Notch intracellular domains *in vivo*. *Cell* **74**, 331–345 (1993).
30. Lawrence, P. A., Johnston, P., Macdonald, P. & Struhl, G. Borders of parasegments in *Drosophila* embryos are delimited by the *fushi tarazu* and *even-skipped* genes. *Nature* **328**, 440–442 (1987).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank M. Fujioka and J. Jaynes for transgenic flies containing the *eve* 4 + 6-*lacZ* construct; L. Andrioli for discussions and support; A. Oberstein for technical assistance; and C. Desplan, J. Blau and T. Cook for encouragement and comments on the manuscript. D.E.C. was supported by a grant from the NSF. This work was also supported by a grant from the NIH.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.S. (stephen.small@nyu.edu).

The gene product Murr1 restricts HIV-1 replication in resting CD4⁺ lymphocytes

Lakshmanan Ganesh¹, Ezra Burstein^{1,2}, Anuradha Guha-Niyogi¹, Mark K. Louder¹, John R. Mascola¹, Leo W. J. Klomp³, Cisca Wijmenga³, Colin S. Duckett² & Gary J. Nabel¹

¹Vaccine Research Center, NIAID, National Institutes of Health, Building 40, Room 4502, MSC-3005, 40 Convent Drive, Bethesda, Maryland 20892-3005, USA

²University of Michigan, Medical Science I, Room 5315, 1301 Catherine Street, Ann Arbor, Michigan 48109-0602, USA

³University Medical Center Utrecht, Lundlaan 6, 3584 EA Utrecht, The Netherlands

Although human immunodeficiency virus-1 (HIV-1) infects quiescent and proliferating CD4⁺ lymphocytes, the virus replicates poorly in resting T cells^{1–6}. Factors that block viral replication in these cells might help to prolong the asymptomatic phase of HIV infection⁷; however, the molecular mechanisms that control this process are not fully understood. Here we show that Murr1, a gene product known previously for its involvement in copper regulation^{8,9}, inhibits HIV-1 growth in unstimulated CD4⁺ T cells. This inhibition was mediated in part through its ability to inhibit basal and cytokine-stimulated nuclear factor (NF)-κB activity. Knockdown of Murr1 increased NF-κB activity and decreased IκB-α concentrations by facilitating phospho-IκB-α degradation by the proteasome. Murr1 was detected in CD4⁺ T cells, and RNA-mediated interference of Murr1 in primary resting CD4⁺ lymphocytes increased HIV-1 replication. Through its effects on the proteasome, Murr1 acts as a genetic restriction factor that inhibits HIV-1 replication in lymphocytes, which

could contribute to the regulation of asymptomatic HIV infection and the progression of AIDS.

Murr1 is a highly conserved 190-amino-acid protein that does not have any identifiable motifs, and a homozygous deletion in the gene encoding canine Murr1 leads to copper toxicosis in Bedlington terriers⁸. In this study, Murr1 was initially identified in a two-hybrid screen by binding the X-linked inhibitor of apoptosis, a known activator of NF-κB (refs 10, 11, and E.B., unpublished observations). To study its effect on NF-κB, HIV-1 reporter plasmids with wild-type or mutant (ΔκB) sites² were co-transfected with control or Murr1 expression plasmids in the different cell lines. Murr1 inhibited both basal and tumour necrosis factor (TNF)-α-dependent HIV-1 transcription from the wild-type but not the κB-mutant reporter in Jurkat T-leukaemia and 293T renal epithelial cell lines (Fig. 1a, left and middle panels). In contrast, Murr1 did not substantially inhibit tumour growth factor-β-dependent transcription in HepG2 cells, confirming its specificity (Fig. 1a, right panel). The κB effect was dose-dependent and observed with other inducers of NF-κB, including interleukin-1 (IL-1) and 12-O-tetradecanoylphorbol-13-acetate (TPA) (Fig. 1b). Murr1 modulated the expression of endogenous κB-regulated genes: transfection into 293T cells decreased the endogenous cell-surface expression of major histocompatibility complex (MHC) class I, in contrast to CD9, which is independent of NF-κB (Fig. 1c).

Its site of action in the NF-κB signalling pathway was further defined by co-transfection of different regulators with an NF-κB reporter in Jurkat T cells. Whereas Murr1 inhibited both IKK-1- and IKK-2-induced NF-κB activity (Fig. 1d, middle and right panels), it failed to block RelA-mediated transcription (Fig. 1d, left panel), indicating that Murr1 might interact downstream of the IκB kinase signalosome. As determined by immunoprecipitation, co-transfected haemagglutinin (HA)-tagged Murr1 and Myc-tagged IKK-2 did not associate *in vivo* (Fig. 2a, lane 2, left panel). Although IKK-1 also did not associate with Murr1 (data not shown), an interaction between transfected HA-tagged Murr1 and endogenous IκB-α was readily detected (Fig. 2a, lane 6). The ankyrin domain of IκB-α was required for association with Murr1, as were amino acids 1–160 of Murr1 (Supplementary Fig. 1a).

A polyclonal antibody against Murr1 demonstrated the association between endogenous Murr1 and IκB-α *in vivo*. RelA antibody immunoprecipitated IκB-α, IκB-β and Murr1 (Fig. 2b, lane 10). IκB-α antibody also pulled down RelA and Murr1 (Fig. 2b, lane 12), but the IκB-β antibody did not precipitate Murr1 (Fig. 2b, lane 14), suggesting that Murr1 interacted preferentially with the NF-κB-IκB-α complex. This association was confirmed *in vivo* by confocal microscopy with fluorescent fusion proteins (Supplementary Fig. 1b), similarly to the pattern of RelA association with IκB-α^{12–14}.

The physiological consequences of these interactions were determined by knockdown of endogenous Murr1 in 293T cells using control and Murr1-specific short interfering RNA (siRNA) duplexes. The specificity of two such siRNAs, Murr1-1 and Murr1-2, directed to different Murr1 sequences, was first confirmed by transfecting 293T cells with wild-type or mutant siRNAs along with wild-type or mutant Murr1 complementary DNAs modified at the siRNA target site (Supplementary Fig. 2). Transient transfection of Murr1-specific siRNA duplexes downregulated endogenous Murr1 and IκB-α, had little effect on IκB-β, p65 or IKK-2 (Fig. 3a, left panel), and increased κB-dependent reporter activity (Fig. 3a, right panel).

To investigate the mechanism of Murr1 action, 293T cells were transfected with a control or Murr1 siRNA. Four days after transfection, cells were treated with the proteasome inhibitor MG132 for 2 h or with vehicle alone and stimulated with TNF-α. Cells depleted of Murr1 showed a decrease in basal IκB-α (Fig. 3a) and an increase and persistence of phospho-IκB-α in response to stimulation with TNF-α (Fig. 3b, left panel). This effect was observed in the absence of a proteasome inhibitor, MG132, but not in its presence (Fig. 3b, right panel), indicating that Murr1

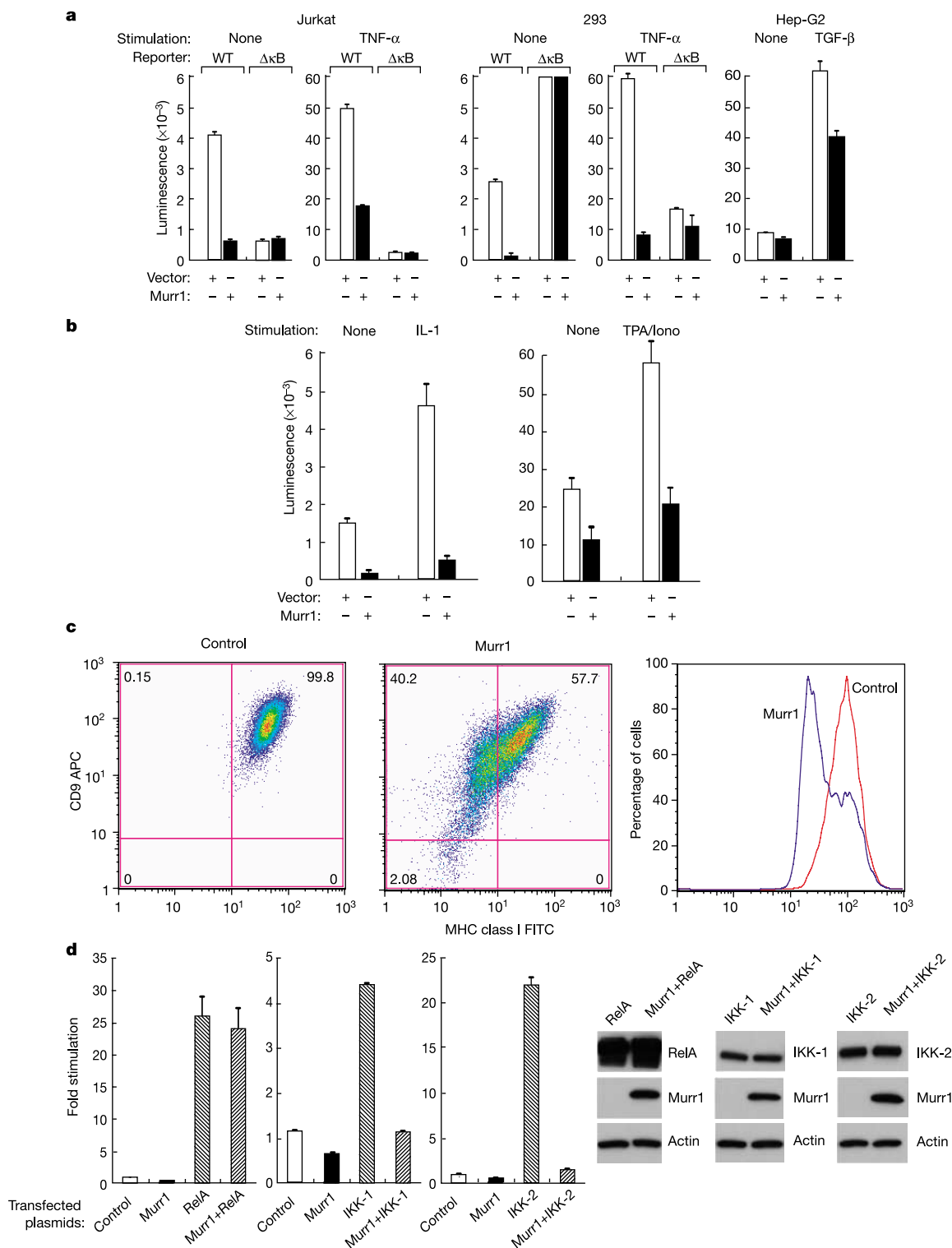


Figure 1 Murr1 suppresses NF- κ B-dependent activity from independent stimuli by acting downstream of κ B kinase. **a**, Murr1 inhibits κ B-dependent gene expression in Jurkat T-leukaemia cells (left) or 293T human embryonic kidney cells (middle) transfected with HIV (WT) or $\Delta\kappa$ B-luciferase reporter ($\Delta\kappa$ B) and Murr1 (solid bars) or vector plasmid (open bars). HepG-2 cells (right) were transfected with p3TP-Lux reporter. Cells were treated with the indicated cytokines after 24 h, and luciferase activity was measured at 36 h. **b**, Murr1 inhibits NF- κ B induced by IL-1 and TPA in 293 cells. **c**, Murr1 decreases the

expression of NF- κ B-dependent endogenous class I MHC in 293T cells transfected with HA-LacZ (left) or HA-Murr1 (middle) by flow cytometry for class I and CD9 in HA-positive cells. The decrease in MHC class I expression is summarized (right). **d**, Murr1 blocks the activation by IKK-1 and IKK-2 but does not inhibit RelA transactivation. Fold stimulation (left) and expression (right) are shown. An average of three independent experiments in triplicate, standardized for transfection efficiency, are shown (**a**, **b**, **d**).

might act similarly to a proteasome inhibitor to enhance and sustain phospho-I κ B. Phospho-I κ B- α is degraded by ubiquitin–protein conjugates that require three enzymes, which participate in a cascade of ubiquitin transfer reactions: a ubiquitin-activating enzyme (E1), a ubiquitin-conjugating enzyme (E2) and a ubiquitin ligase (E3). The specificity of protein ubiquitination and 26S proteasome degradation is determined by the E3 enzymes¹⁵. Skp1/Cul1/ β -TrCP1 (F-Box) complexes constitute a class of E3 enzymes that are required for the degradation of phospho-I κ B- α ¹⁶. The possible association of endogenous Murr1 with Skp1/Cul1/ β -TrCP1 (F-box) complexes was examined. As determined by immunoprecipitation followed by western blotting, Murr1 interacted biochemically with the Cul1 and not the Skp1 component of the E3 ligase complex, in contrast to a denatured negative control (Fig. 3c). Although Murr1 binds to common constituents of E3 ligase, it nevertheless showed specificity because knockdown did not alter steady-state concentrations of β -catenin, the target of proteasomal degradation of the Wnt signalling pathway (Fig. 3d). Taken together, these results suggest that Murr1 blocked NF- κ B activation through its ability to interact with E3 ligase and inhibit the proteasomal degradation of I κ B.

To determine whether Murr1 was detectable in T cells, lymphocytes from normal healthy individuals were examined. Both CD4⁺ and CD8⁺ cells expressed Murr1, with higher concentrations in CD4⁺ cells (Fig. 4a), raising the possibility that Murr1 might affect HIV-1 replication in these cells. This possibility was examined first by transfection of Murr1 in primary T cells and compared with another inhibitor of NF- κ B activation, the I κ B super-repressor (I κ B-SR). Expression of recombinant Murr1 or I κ B-SR inhibited HIV-1 replication in activated T cells similarly relative to controls (Fig. 4b; $P = 0.02$; paired t -test), indicating that Murr1 might inhibit HIV-1 replication comparably to NF- κ B inhibitors in primary activated T cells. To determine its effect on HIV-1 replication in resting primary T cells, CD4⁺ cells from HIV-1-negative donors were transfected with an siRNA for Murr1 (Murr1-1) or a negative control. To identify transfected cells, siRNAs were mixed with an unrelated fluorescent Cy3-labelled luciferase siRNA at a ratio of 2:1, and more than 30% of cells displayed fluorescence. In Murr1 siRNA-transfected cells, a maximum decrease in Murr1

protein concentration was achieved between 48 and 72 h after electroporation, and these cells did not display T-cell activation markers—CD69, CD25 and HLA-DR—at the time of HIV-1 infection (data not shown). Murr1 siRNA increased dose-dependent HIV-1 replication in primary resting CD4⁺ T cells from three separate cell donors (Fig. 4c; $P = 0.0007$, 0.0024 and 0.0014, respectively; paired t -test), implicating Murr1 in the regulation of HIV infection in these cells.

Whereas HIV-1 entry into activated CD4⁺ lymphocytes leads to a productive infection¹, the virus remains latent in resting CD4⁺ lymphocytes¹⁷, existing as a preintegration complex awaiting cell stimulation, which facilitates the transition to productive replication^{3,4,6,7,18}. The molecular mechanisms responsible for HIV-1

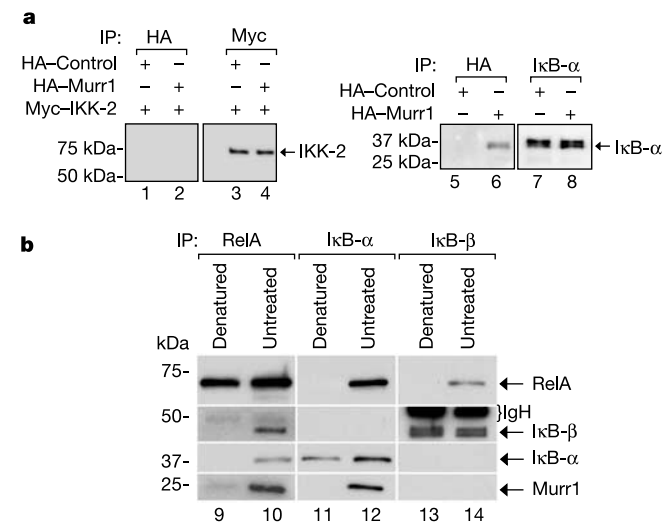


Figure 2 Murr1 associates with the NF- κ B-I κ B- α complex. **a**, Murr1 does not associate with IKK-2 but interacts with endogenous I κ B- α . Lysates from transfected 293T cells, treated with lactacystin (10 μ M) 16 h before harvesting, were immunoprecipitated (IP) as indicated and immunoblotted with antibody against IKK-2 (left) or I κ B- α (right). **b**, Immunoprecipitation of Murr1–NF- κ B–I κ B- α in 293T cells with antibodies against indicated proteins. IgH, immunoglobulin heavy chain.

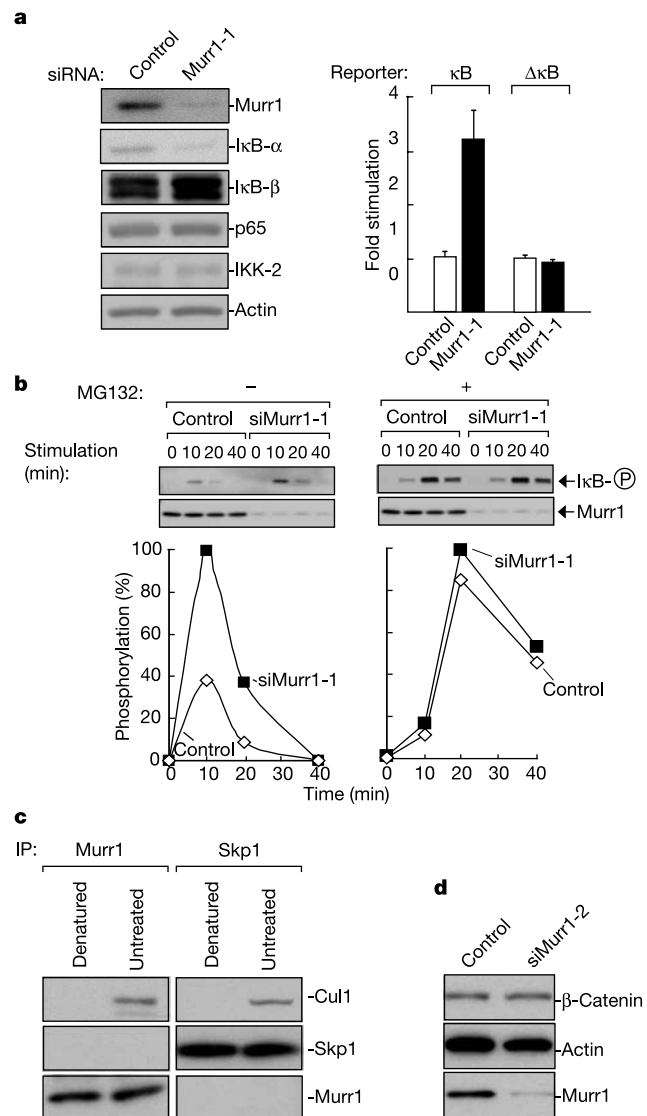


Figure 3 Murr1 regulates I κ B- α turnover through effects on proteasomal degradation. **a**, Knockdown of Murr1 decreases cellular I κ B- α and increases basal NF- κ B activity in 293T cells transfected with control or Murr1 siRNAs, analysed by immunoblotting to the indicated proteins at 48 h (left). Fold stimulation of the indicated co-transfected plasmids is shown (right). **b**, Murr1 alters I κ B- α phosphorylation in 293T cells transfected with control or Murr1 siRNAs. At 90 h after transfection, cells were treated with vehicle (left) or MG132 (right) for 2 h, followed by stimulation with TNF- α and immunoblotting for Murr1 and phospho-I κ B- α (top) and quantification (bottom). **c**, Immunoprecipitation (IP) of Skp1–Cul1– β -TrCP1 complexes in 293T cell extracts with antibodies against Murr1 or Skp1, immunoblotted for Cul1, Skp1 or Murr1. **d**, Knockdown of Murr1 by Murr1-2 siRNA compared with a negative control (GFP) does not alter β -catenin concentrations.

latency are not well understood. Among the factors that can activate HIV-1 replication are host transcription factors such as NF- κ B (ref. 2), epigenetic modifications at the site of integration¹⁹, premature termination of transcription due to the absence of sufficient concentrations of Tat and NF- κ B (refs 20, 21), and inefficient export of RNAs for structural proteins⁵. The roles of these factors in resting CD4⁺ T cells are inferred from cell culture experiments and affect viral transcription, translation or RNA transport. Here we show that Murr1 regulates I κ B- α turnover, which inhibits the productive HIV-1 infection of resting T lymphocytes. Expression of Murr1 in activated HIV-infected T cells also inhibited virus replication, indicating that it might also have a function in limiting the extent of viral replication, or burst size. NF- κ B is sequestered in the cytoplasm by a 450-kDa I κ B complex^{12,22–25} and, after activation of T cells, phosphorylation of I κ B- α (ref. 26) in this complex by two

inducible kinases, IKK-1 and IKK-2, leads to the subsequent ubiquitination and degradation of I κ B- α by the 26S proteasome²⁷. Degradation of I κ B- α by the proteasome releases NF- κ B, which then translocates to the nucleus to stimulate the synthesis of genes involved in immune activation. NF- κ B enhances transcription and subsequently recruits several transcriptional activators, including positive transcription elongation factor b, which is ubiquitinated and degraded by the proteasome²⁸. Inhibition of proteasomal activity by Murr1, after it rebounds from its initial degradation and after this processive transcription complex has formed, could subsequently stabilize it and enhance transcription. The proteasome, through the ubiquitin E3 ligase complex, is also involved in viral processing and the maturation of Gag²⁹. Because Murr1 inhibits proteasomal degradation, it could further block HIV-1 infection by decreasing the proteasome-dependent degradation and maturation of Gag-associated factors required for viral replication, and although Murr1 acts selectively on the NF- κ B and not the Wnt signalling pathway (Fig. 3), it remains possible that it might yet affect other such proteins degraded by the proteasome. Such an example is APOBEC3G, shown recently to associate with a Cul5 E3 ligase³¹. Blocks to viral replication in resting lymphocytes, as well as in cells containing latent provirus, might affect disease progression and viral rebound after the discontinuation of anti-retroviral therapy (reviewed in ref. 7). A more precise understanding of the molecular regulation of these events might identify novel genetic restriction factors, like Murr1, through which anti-viral drugs might delay the progression of HIV-1 infection to AIDS. □

Methods

Plasmids, siRNA and cell transfections

The gene encoding chloramphenicol acetyltransferase (CAT) in HIV-LTR-CAT and HIV-LTR- Δ IKB-CAT² was replaced with the luciferase gene from pGL3 basic (Promega) to construct HIV-LTR-Luc and HIV-LTR- Δ IKB-Luc. The HA-I κ B- α SR, p3TP-Lux reporter, green fluorescent protein (GFP)-RelA, pRK-Myc-IKK-1, pRK-Myc-IKK-2, p65 and I κ B- α cDNAs are described in Supplementary Methods, as are siRNA sequences and the cell transfection techniques.

T-cell purification and transfection

To purify resting T cells, CD4⁺ T lymphocytes were isolated from Ficoll-purified peripheral blood mononuclear cells (PBMCs) by negative selection with a CD4⁺ T-cell isolation kit containing CD8, CD11b, CD16, CD19, CD36 and CD56 antibodies (Miltenyi Biotec). CD25⁺ and HLA-DR⁺ cells were then removed from this population by using CD69, CD25 and HLA-DR antibodies (Miltenyi Biotec) by negative selection. The purity of the isolated resting T cells was assessed by fluorescence-activated cell sorting (FACS) analysis for CD4, CD69, CD25 and HLA-DR (BD-Pharmingen); 95% of the cells were CD4⁺ and there were less than 0.1% of activated T cells as defined by the presence of CD69, CD25 or HLA-DR. Resting CD4⁺ T cells (10^7) were transfected with control or Murr1 siRNA (Dharmacon). For each transfection, 1.2 μ M siRNA (800 nM unlabelled siRNA and 400 nM Cy3-labelled siRNA) was used. Transfections were performed with a 2:1 mixture of unlabelled siRNA and Cy3 luciferase. To transfect the resting CD4⁺ T cells with siRNA, 10^7 cells were resuspended in 100 μ l T-cell Nucleofector reagent and immediately electroporated with the recommended protocol U-14 on the Nucleofector instrument (Amax Biosystems). The electroporated cells were washed 4 h after transfection and incubated in 1.5 ml RPMI and infected with HIV-1 48 h after electroporation. FACS analysis was performed with CD69, CD25 and HLA-DR antibodies (BD Pharmingen) to check the activation status of the cells at the time of infection. For plasmid transfections, CD4⁺ T lymphocytes isolated from Ficoll-purified PBMCs by negative selection were stimulated with phytohemagglutinin (10μ g ml⁻¹) (Calbiochem) and 40 U ml⁻¹ IL-2 (Peprotech) for 24 h and maintained in 40 U ml⁻¹ IL-2 for 4 days. Cells were then washed to remove IL-2, and 10^7 cells were resuspended in 100 μ l T-cell Nucleofector reagent containing 0.5 μ g of the respective plasmids and immediately electroporated with recommended protocol T-20 on the Nucleofector instrument. The electroporated cells were washed 12 h after transfection and incubated in 1.5 ml of RPMI supplemented with 40 U ml⁻¹ IL-2 and infected with HIV-1 24 h after electroporation.

HIV-1 infection and flow cytometric analysis for expression of p24-Gag

At 48 h after siRNA transfection or 24 h after plasmid transfection, HIV-1 infection of CD4⁺ T lymphocytes was performed in 96-well round-bottomed culture plates by combining 40 μ l virus stock with 20 μ l CD4⁺ T lymphocytes (1.5×10^5 cells). The multiplicity of infection of undiluted virus was ~ 1.0 . After incubation for 2 h at 37 °C, cells were washed twice and continued in culture with indinavir at 1 μ M. CD4⁺ T lymphocytes were harvested for intracellular p24-Gag staining 48 h after exposure to virus. In brief, cells were stained with anti-CD4-PE and ethidium bromide monoazide (EMA) for 10 min. EMA was cross-linked to the cells by exposure to a bright light source for 15 min. Cells were washed once, fixed and permeabilized with Cytoperm/Cytofix (BD Pharmingen) for

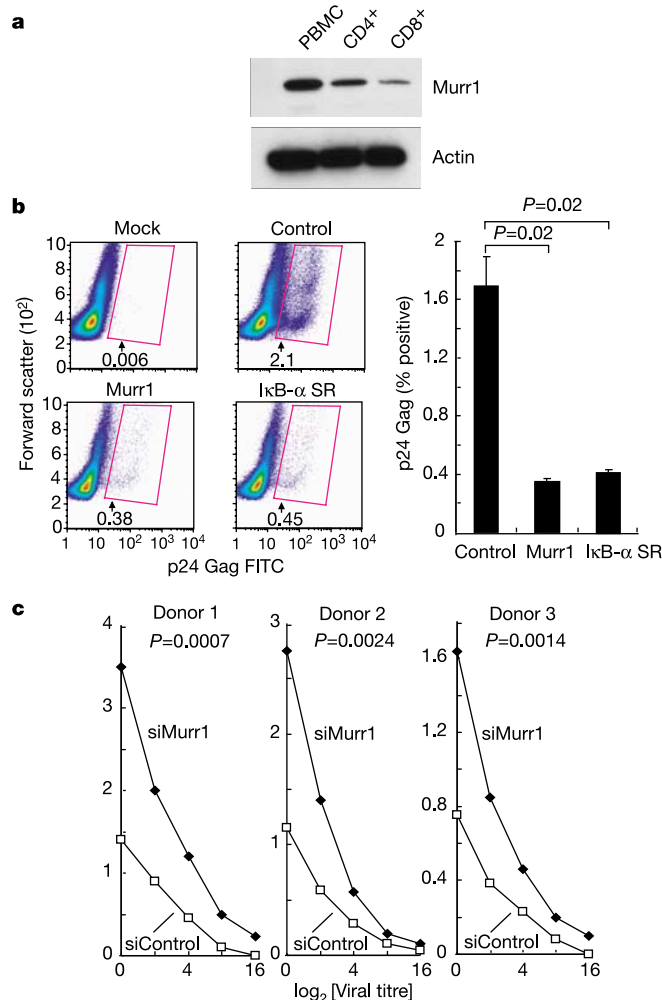


Figure 4 Murr1 inhibits productive HIV-1 infection in CD4⁺ lymphocytes. **a**, Murr1 concentrations in CD4⁺ and CD8⁺ T cells shown by immunoblotting with Murr1 antisera and reprobing with β -actin antibody. **b**, Murr1 inhibits HIV-1 replication in phytohemagglutinin/IL-2-stimulated human CD4⁺ cells transfected with plasmids encoding HA-LacZ (control), HA-Murr1 or HA-I κ B-SR, infected with HIV-1_{BAL} (p24 = 75 ng ml⁻¹) 36 h later. HIV-1 replication was analysed 48 h after infection by staining for intracellular p24. A representative of three independent HIV-1-negative donors is shown ($P = 0.02$, control versus Murr1). **c**, Murr1 siRNA-transfected resting CD4⁺ T lymphocytes from three healthy donors show increased susceptibility to HIV-1 infection after transfection with siRNA for Murr1 (Murr1-1) compared with GFP (control). Cy3 luciferase siRNA was mixed with siRNAs (2:1) to label transfected cells. Single-round HIV-1 replication was measured in labelled cells at 48 h. The multiplicity of infection of undiluted virus was ~ 1.0 .

20 min. Cells were then stained for p24 Gag (KC-57 fluorescein isothiocyanate (FITC); Coulter) for 20 min and washed once in 1 × Perm/wash buffer (BD Pharmingen). Cells in each group were analysed by flow cytometry for intracellular HIV-1 p24 and CD4 after the exclusion of dead cells by their affinity for EMA. For a detailed description of HIV-1 production, infection and flow cytometric analysis see ref. 30.

Antibodies, western blots and immunoprecipitations

Cell lysis for western blots and immunoprecipitations were performed in cell lysis buffer (Cell Signal). Antibodies against β -actin (Sigma), His (Invitrogen), HA, HA-PE, p65, I κ B- α , I κ B- β , Cul1, Skp1 (Santa Cruz), IKK1, IKK2, Ser-32-phospho I κ B- α , β -catenin (Cell Signal) and HLA ClassI FITC (Biosource) were used in flow cytometry, immunoblotting and immunoprecipitations in accordance with the manufacturer's instructions. Polyclonal Murr1 antibody has been described previously⁹.

Received 11 August; accepted 21 October 2003; doi:10.1038/nature02171.

- McDougal, J. S. *et al.* Cellular tropism of the human retrovirus HTLV-III/LAV 1. Role of T cell activation and expression of the T4 antigen. *J. Immunol.* **135**, 3151–3162 (1985).
- Nabel, G. & Baltimore, D. An inducible transcription factor activates expression of human immunodeficiency virus in T cells. *Nature* **326**, 711–713 (1987).
- Zack, J. A. *et al.* HIV-1 entry into quiescent primary lymphocytes: molecular analysis reveals a labile, latent viral structure. *Cell* **61**, 213–222 (1990).
- Bukrinsky, M. I., Stanwick, T. L., Dempsey, M. P. & Stevenson, M. Quiescent T lymphocytes as an inducible virus reservoir in HIV-1 infection. *Science* **254**, 423–427 (1991).
- Pomerantz, R. J., Seshamma, T. & Trono, D. Efficient replication of human immunodeficiency virus type 1 requires a threshold concentration of Rev: potential implications for latency. *J. Virol.* **66**, 1809–1813 (1992).
- Zack, J. A., Haislip, A. M., Krogstad, P. & Chen, I. S. Incompletely reverse-transcribed human immunodeficiency virus type 1 genomes in quiescent cells can function as intermediates in the retroviral life cycle. *J. Virol.* **66**, 1717–1725 (1992).
- Blankson, J. N., Persaud, D. & Siliciano, R. F. The challenge of viral reservoirs in HIV-1 infection. *Annu. Rev. Med.* **53**, 557–593 (2002).
- van de Sluis, B., Rothuizen, J., Pearson, P. L., van Oost, B. A. & Wijmenga, C. Identification of a new copper metabolism gene by positional cloning in a purebred dog population. *Hum. Mol. Genet.* **11**, 165–173 (2002).
- Klomp, A. E. M., van de Sluis, B., Klomp, L. W. J. & Wijmenga, C. The ubiquitously expressed Murr1 protein is absent in canine copper toxicosis. *J. Hepatol.* **39**, 703–709 (2003).
- Hofer-Warbinek, R. *et al.* Activation of NF- κ B by XIAP, the X chromosome-linked inhibitor of apoptosis, in endothelial cells involves TAK1. *J. Biol. Chem.* **275**, 22064–22068 (2000).
- Richter, B. W. M. & Duckett, C. S. The IAP proteins: caspase inhibitors and beyond. *Sci. STKE* <http://stke.sciencemag.org/cgi/content/full/sigtrans2000/44/pe1> (2000).
- Verma, I. M., Stevenson, J. K., Schwarz, E. M., Antwerp, D. V. & Miyamoto, S. Rel/NF- κ B/I κ B family: intimate tales of association and dissociation. *Genes Dev.* **9**, 2723–2735 (1995).
- Huxford, T., Huang, D.-B., Malek, S. & Ghosh, G. The crystal structure of the I κ B α /NF- κ B complex reveals mechanisms of NF- κ B inactivation. *Cell* **95**, 759–770 (1998).
- Schmid, J. A. *et al.* Dynamics of NF κ B and I κ B α studied with green fluorescent protein (GFP) fusion proteins. Investigation of GFP-p65 binding to DNA by fluorescence resonance energy transfer. *J. Biol. Chem.* **275**, 17035–17042 (2000).
- Hersko, A. & Ciechanover, A. The ubiquitin system. *Annu. Rev. Biochem.* **67**, 425–479 (1998).
- Tan, P. *et al.* Recruitment of a ROC1-CUL1 ubiquitin ligase by Skp1 and HOS to catalyze the ubiquitination of I κ B α . *Mol. Cell* **3**, 527–533 (1999).
- Wong, J. K. *et al.* Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. *Science* **278**, 1291–1295 (1997).
- Finzi, D. *et al.* Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science* **278**, 1295–1300 (1997).
- Jordan, A., Defechereux, P. & Verdin, E. The site of HIV-1 integration in the human genome determines basal transcriptional activity and response to Tat transactivation. *EMBO J.* **20**, 1726–1738 (2001).
- Kao, S. Y., Calman, A. F., Luciw, P. A. & Peterlin, B. M. Anti-termination of transcription within the long terminal repeat of HIV-1 by tat gene product. *Nature* **330**, 489–493 (1987).
- Adams, M. *et al.* Cellular latency in human immunodeficiency virus-infected individuals with high CD4 levels can be detected by the presence of promoter-proximal transcripts. *Proc. Natl Acad. Sci. USA* **91**, 3862–3866 (1994).
- Baldwin, A. S. Jr The transcription factor NF- κ B and human disease. *J. Clin. Invest.* **107**, 3–6 (2001).
- Silverman, N. & Maniatis, T. NF- κ B signaling pathways in mammalian and insect innate immunity. *Genes Dev.* **15**, 2321–2342 (2001).
- Ben-Neriah, Y. Regulatory functions of ubiquitination in the immune system. *Nature Immunol.* **3**, 20–26 (2002).
- Karin, M. & Lin, A. NF- κ B at the crossroads of life and death. *Nature Immunol.* **3**, 221–227 (2002).
- Beg, A. A., Finco, T. S., Nantermet, P. V. & Baldwin, A. S. Jr Tumor necrosis factor and interleukin-1 lead to phosphorylation and loss of I κ B- α : a mechanism for NF- κ B activation. *Mol. Cell. Biol.* **13**, 3301–3310 (1993).
- Zandi, E., Rothwarf, D. M., Delhase, M., Hayakawa, M. & Karin, M. The I κ B kinase complex (IKK) contains two kinase subunits, IKK α and IKK β , necessary for I κ B phosphorylation and NF- κ B activation. *Cell* **91**, 243–252 (1997).
- Kiernan, R. E. *et al.* Interaction between cyclin T1 and SCF^{SKP2} targets CDK9 for ubiquitination and degradation by the proteasome. *Mol. Cell. Biol.* **21**, 7956–7970 (2001).
- Schubert, U. *et al.* Proteasome inhibition interferes with Gag polyprotein processing, release, and maturation of HIV-1 and HIV-2. *Proc. Natl Acad. Sci. USA* **97**, 13057–13062 (2000).
- Mascola, J. R. *et al.* Human immunodeficiency virus type 1 neutralization measured by flow cytometric quantitation of single-round infection of primary human T cells. *J. Virol.* **76**, 4810–4821 (2002).
- Yu, X. *et al.* Induction of APOBEC3G ubiquitination and degradation by an HIV-1 Vif-Cul5-SCF complex. *Science* **302**, 1056–1060 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank K. Leung for preparation of the Murr1 mutants; S. Majeed for helping with protein purifications; M. Roederer for advice in interpretation of flow cytometry data and statistical analysis; A. Tislerics, A. Gooch and T. Suhana for manuscript preparation; and K. Stroud and T. Miller for the preparation of figures. This work was supported in part by the Michigan Gastrointestinal Hormone Research Core Center (E.B.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.J.N. (gnabel@nih.gov).

Inheritance of a pre-inactivated paternal X chromosome in early mouse embryos

Khanh D. Huynh & Jeannie T. Lee

Howard Hughes Medical Institute, Department of Molecular Biology, Massachusetts General Hospital, Department of Genetics, Harvard Medical School, Boston, Massachusetts 02114, USA

In mammals, dosage compensation ensures equal X-chromosome expression between males (XY) and females (XX) by transcriptionally silencing one X chromosome in XX embryos¹. In the prevailing view, the XX zygote inherits two active X chromosomes, one each from the mother and father, and X inactivation does not occur until after implantation^{2–6}. Here, we report evidence to the contrary in mice. We find that one X chromosome is already silent at zygotic gene activation (2-cell stage). This X chromosome is paternal in origin and exhibits a gradient of silencing. Genes close to the X-inactivation centre show the greatest degree of inactivation, whereas more distal genes show variable inactivation and can partially escape silencing. After implantation, imprinted silencing in extraembryonic tissues becomes globalized and more complete on a gene-by-gene basis. These results argue that the XX embryo is in fact dosage compensated at conception along much of the X chromosome. We propose that imprinted X inactivation results from inheritance of a pre-inactivated X chromosome from the paternal germ line.

In mice, X-chromosome inactivation (XCI) takes on two lineage-specific forms. Random XCI¹, in which both X chromosomes have an equal chance of being inactivated, occurs in the epiblast (embryo proper). In contrast, imprinted XCI⁷ leads to paternal X-chromosome (X^P) silencing in the extraembryonic tissues (placenta). Although the imprint is set in gametes, classical studies support a view in which both X^P and X^M (maternal X chromosome) are transmitted to the zygote in an active form: the absence of a late-replicating X chromosome in pre-implantation embryos² suggests equal X-transcriptional status, and bimodal distributions of enzymatic activities for two X-linked genes (*Hprt*^{3,4}, *Gla*⁵) suggest twice as much expression in XX as compared with XY embryos. Thus, the prevailing view postulates that XX and XY embryos have a twofold imbalance of X dosage until implantation, when XCI takes place for the first time in the extraembryonic and embryonic lineages^{6,8–10}. However, some recent observations have not been explained easily¹¹. Studies of the X-linked *Pgk1* (refs 12, 13) revealed that, whereas the maternal allele is expressed during pre-implantation development, the paternal allele is silent. Furthermore, one *Xist* allele is expressed at high levels in pre-implantation embryos⁹, leading to the idea that the X chromosomes may be transcriptionally distinct^{6,11}. Indeed, the maternal and paternal haplogenomes can behave differently in early

development¹⁴. To distinguish between the opposing views, we now re-visit the question of pre-implantation X-chromosome activity using techniques unavailable in the 1970s.

First, we determined whether there was large-scale X-linked transcription in pre-implantation embryos using a recently developed RNA fluorescence *in situ* hybridization (FISH) technique in which a Cot-1 probe is used to detect nascent RNA in non-denatured nuclei¹⁵ (Fig. 1). Because the Cot-1 fraction is enriched for repetitive elements that frequently occur in introns and 3' untranslated regions, Cot-1 RNA signals mark regions of ongoing transcription. An actively transcribing Xist-coated X chromosome would therefore show coincident Cot-1 signals. We collected XX embryos from the 1-cell zygote to morula and performed RNA-FISH using a Cot-1 (red) and an Xist (green) probe. None of the 1-cell zygotes showed Cot-1 signals (Fig. 1a), consistent with a transcriptionally dormant state.

Cot-1 signals became apparent at the 2-cell stage (Fig. 1b) when zygotic gene activation (ZGA) occurs for the first time (reviewed in ref. 16). In 2-cell embryos, Cot-1 signals were not evident in all cells,

suggesting that ZGA had not uniformly occurred (Fig. 1g). In contrast, zygotic transcription was seen in all blastomeres of other cleavage stages (Fig. 1c–g). Notably, the Xist RNA compartment was devoid of Cot-1 staining, whereas the remaining nucleoplasm exhibited diffuse Cot-1 hybridization (Fig. 1b–e; note that single z-sections reveal events on only a single plane). This Xist-coincident Cot-1 'hole' was observed in 83–100% of all Xist-positive cells during all cleavage stages, and suggested inactivity of Xist-associated domains. Negative staining of the transcriptionally inert polar body (asterisk) in each embryo indicated that the Cot-1 probe was indeed detecting RNA and not DNA. In addition to Xist compartments, Cot-1 staining was also absent in presumptive pericentric heterochromatin and nucleoli. Collectively, these results indicate that one X chromosome is transcription-poor throughout pre-implantation female development.

The appearance of Xist RNA was highly variable among Cot-1-positive blastomeres of the same stage. Although an approximately 80% majority of blastomeres exhibited extensive Xist RNA coating, a significant minority showed either very small Xist signals or none

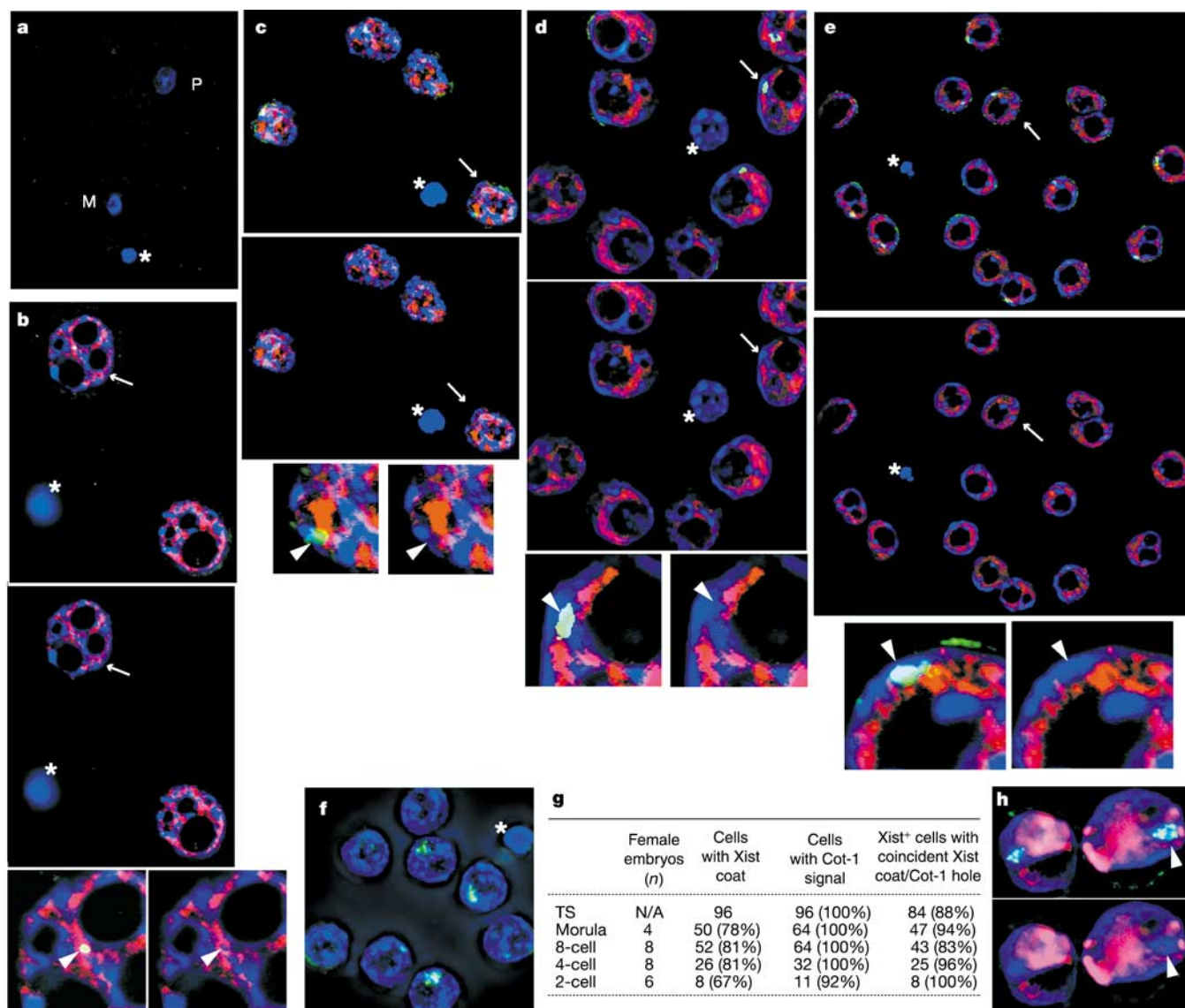


Figure 1 Paucity of transcription on one X chromosome in pre-implantation embryos. RNA-FISH analysis with Texas red-labelled mouse Cot-1 (red) and Alexa Fluor-488-labelled Xist RNA (green) probes. DAPI (blue) stains DNA. Because all images are single z-sections (except where noted), only signals on the same plane are visible. Asterisk, polar body. Arrows indicate nuclei that are enlarged below; arrowheads denote the Cot-1 hole

and Xist compartment. **a**, One-cell zygote. Maternal (M) and paternal (P) pronuclei are evident. **b**, Two-cell embryo. **c**, Four-cell embryo. **d**, Eight-cell embryo. **e**, Sixteen-cell embryo. **f**, Variable Xist expression among blastomeres of an 8-cell embryo. To capture Xist signals in all planes, a two-dimensional projection of 50 0.2- μ m z-sections is shown. **g**, Summary of FISH data. **h**, FISH of TS cells.

at all (Fig. 1f, g). The presence of Cot-1 signals in Xist-negative cells argued against poor probe penetration as a cause of absent Xist signals. This observation is consistent with variable Xist RNA levels in a previous single-cell polymerase chain reaction with reverse transcription (RT-PCR) analysis¹⁷, and may suggest physiological distinctiveness among the blastomeres (see below).

To test whether a Cot-1 hole could be observed in tissues with a well-accepted inactive X chromosome (Xi), we examined XX mouse trophoblast stem (TS) cells¹⁸. Indeed, a Cot-1 hole was also observed in the Xist RNA compartment (Fig. 1g, h), validating the technique for assessing transcriptional inactivity. The Xist domain of TS cells often showed intense 4,6-diamidino-2-phenylindole (DAPI) staining, a hallmark of heterochromatin observed on the human Barr body (Xi), but was not seen in pre-implantation mouse embryos (data not shown). This result suggested that the X^P in TS cells and pre-implantation embryos might not be equivalent (see below).

To determine the origin and extent of XCI, we assessed the transcriptional status of the X^M and X^P on a gene-by-gene basis. Hybrid female embryos were generated by crossing *Mus musculus* (C57BL/6) to *Mus castaneus* (CAST/Ei), strains that carry numerous single-nucleotide polymorphisms (SNPs) that enable us to distinguish between the two X chromosomes. We identified 13 X-linked genes containing SNPs within restriction enzyme sites and carried out allele-specific RT-PCR assays on single morulae (8- to 16-cell stage). Earlier cleavage stages were not analysed in order to avoid RNA that is maternally loaded into the oocyte (see below). Such RNA is generally destroyed at ZGA (2-cell stage) when maternal and paternal haplogenomes are activated^{16,19}.

When CAST/Ei males were mated to C57BL/6 females, several X-linked genes deviated from the expected pattern of equal maternal and paternal expression (Fig. 2a). Most notably, heavily skewed maternal expression was observed in genes within 10 cM of the X-inactivation centre (Xic)²⁰, including *Chic1*, *Xnp*, *Atp7a*, *Pgk1* and *Rlim*. Preferential maternal expression was also observed in

genes within 25 cM in either direction from the Xic, such as *Gla*, *Hprt*, *G6pd*, *Fmr1*, *Abc7* and *Mecp2*. In *Gla*, skewed expression was seen in only a subset of embryos. In contrast, biased expression was not observed in any embryo for *Pctk1* and *Pdha1*, genes far distal to the Xic. Thus, there appeared to be a gradient of silencing, with the highest degree of silencing within 10 cM of the Xic, variable silencing up to 25 cM of the Xic, and escape from silencing near the chromosome ends.

A priori, skewed expression might arise from either strain-specific or parent-of-origin effects (imprinting). To distinguish between them, we carried out the reciprocal cross (Fig. 2b). When C57BL/6 males were crossed with CAST/Ei females, maternal-specific expression was recapitulated in genes within 10 cM of the Xic (*Chic1*, *Xnp*, *Atp7a*, *Pgk1* and *Rlim*). More distal genes also demonstrated preferential maternal expression (*Hprt*, *Mecp2*, *G6pd*, *Abc7* and *Gla*). *Gla* exhibited the same embryo-to-embryo variability, with some showing skewing and others showing equal bi-allelic expression. The only exception to reciprocity was *Fmr1*, which seemed to exhibit both an imprinting effect and a strain-specific preference for the C57BL/6 allele. The most distal genes, *Pctk1* and *Pdha1*, remained exempt from silencing. Thus, the bias on the X chromosome arose from parent-of-origin effects rather than strain-specific effects.

To rule out a general maternal effect, we examined five randomly chosen expressed autosomal genes and found no evidence of persistent maternal RNA in the morula (Fig. 3a). Persistent transcripts would have been manifested by greater contribution of maternal bands in both forward and reverse crosses. However, none of the five genes exhibited a maternal-specific bias, as *B2m*, *Yy1*, *Wars*, *Hras1* and *Nxf1* were all bi-allelically expressed (a mild strain-specific preference was observed for *Wars*, *Hras1* and *Nxf1*). These results suggested that maternal transcripts do not generally persist into the morula stage. This is further supported by equal bi-allelic expression of *Pdha1* and *Pctk1* (Fig. 2). Thus, we concluded

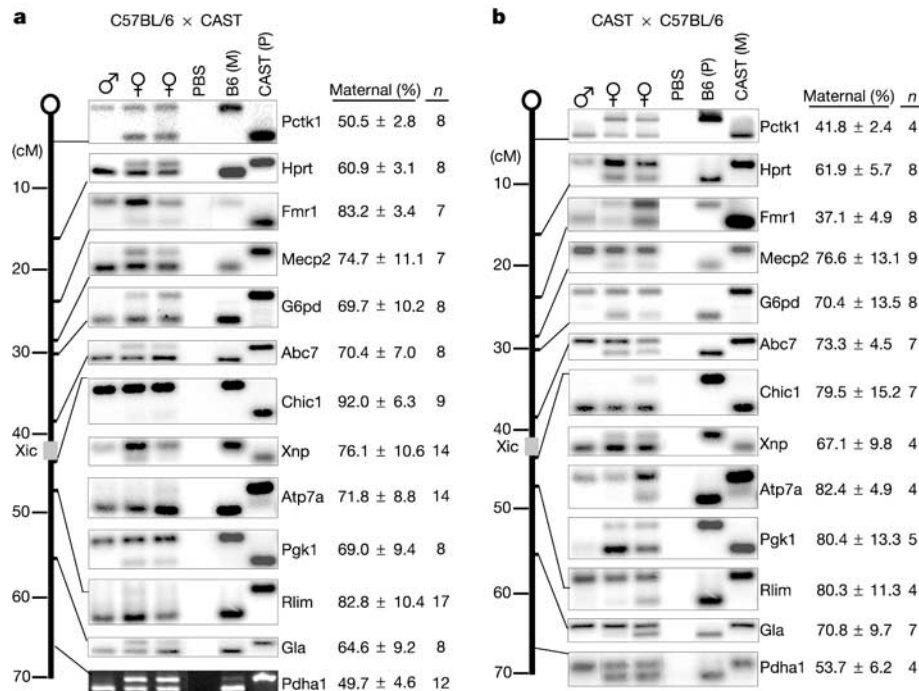


Figure 2 Partial X^P inactivation in pre-implantation embryos. **a, b**, Allele-specific RT-PCR analysis of single morulae from reciprocal crosses between C57BL/6 (B6) and CAST/Ei mice, as indicated (maternal genotype denoted first, by convention). Embryos were sexed by Xist expression. Representative males and females are shown. To exclude maternal contamination, PBS wash sample taken from the last rinse was analysed in

parallel. Data presented are Southern autoradiographic images except for *Pdha1* in **a**, which is from ethidium-bromide-stained gels. The X^M contribution (%) was quantified by hybridization to internal primers (except *Pdha1*, which was quantified by ethidium bromide intensity). Multiple embryos were examined (*n*) and the average X^M contribution is shown (% maternal). s.d., standard deviation; M, maternal; P, paternal.

that preferential X^M expression is zygotically in origin, specific to the X chromosome and is the result of imprinting rather than maternal or strain-specific effects.

Taken together, FISH and allele-specific RT-PCR analyses indicate that the XX pre-implantation mouse embryo is largely dosage compensated from the time of conception. This mechanism of inactivation, however, is clearly imperfect in two respects. First, the X^P exhibits a gradient of silencing, with the strength of silencing highest near *Xic* and frequent escape in distal regions. Second, for any gene subject to inactivation, expression was variable among blastomeres of a single embryo (for example, *Xist*) and among different embryos. Thus, the average X^M contribution is not 100%. It is noteworthy that about 20% of blastomeres on average show markedly diminished or absent *Xist* coats (Fig. 1f, g). One intriguing possibility is that these blastomeres may be precursors of the future epiblast, where imprinting is known to be erased. A 'weakened' *Xist* imprint may partly explain the leakiness of X^P silencing in pre-implantation embryos. Specifically, if one blastomere out of four escapes imprinting, the X^P would hypothetically contribute 20% of total X transcription in that embryo. This incompleteness of imprinting would be consistent with viability of the occasional X^P O embryo, as indeed the loss of X^P silencing would allow some

blastomeres to survive despite no contribution from X^M (reviewed in ref. 11).

We next addressed whether the X^P is remodelled after implantation, a time when imprinted XCI has been postulated to take place for the first time. We examined TS cells, placental precursors that show imprinted X^P silencing¹⁸. Here, we derived two hybrid TS lines from F1(C57BL/6 $\times$ CAST/Ei) blastocysts and showed that they stably carried two X chromosomes (Fig. 3b). Notably, the X-linked genes demonstrated exclusive X^M expression (Fig. 3c): not only was the spread of silencing more extensive along the TS X chromosome, but silencing was also more complete for any given gene. In parallel, we tested their expression patterns in epiblast-derived embryonic stem (ES) cells (carrying one *M. castaneus* and one *M. musculus* X chromosome) and found essentially equal bi-allelic expression (Fig. 3d). These results established that the tested genes are

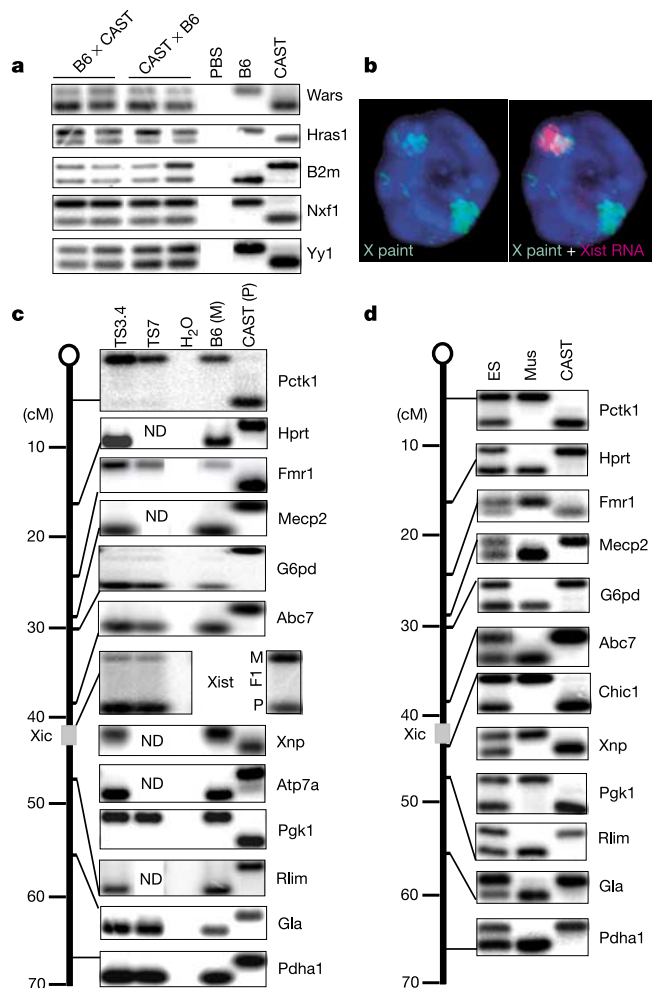


Figure 3 Autosomal expression in pre-implantation embryos and pattern of X-linked expression in post-implantation cells. **a**, Allele-specific RT-PCR analysis of five randomly chosen autosomal genes. **b**, *Xist* RNA-FISH (red) and X-chromosome-specific paint (green) of TS3.4. **c**, Allele-specific RT-PCR of TS cells. TS3.4 is monoclonally derived and TS7 is polyclonally derived from a C57BL/6 $\times$ CAST/Ei cross. **d**, Allele-specific RT-PCR of 16.7 ES cells. Mus, *M. musculus*; F1, control F1 fibroblast line; ND, not done.

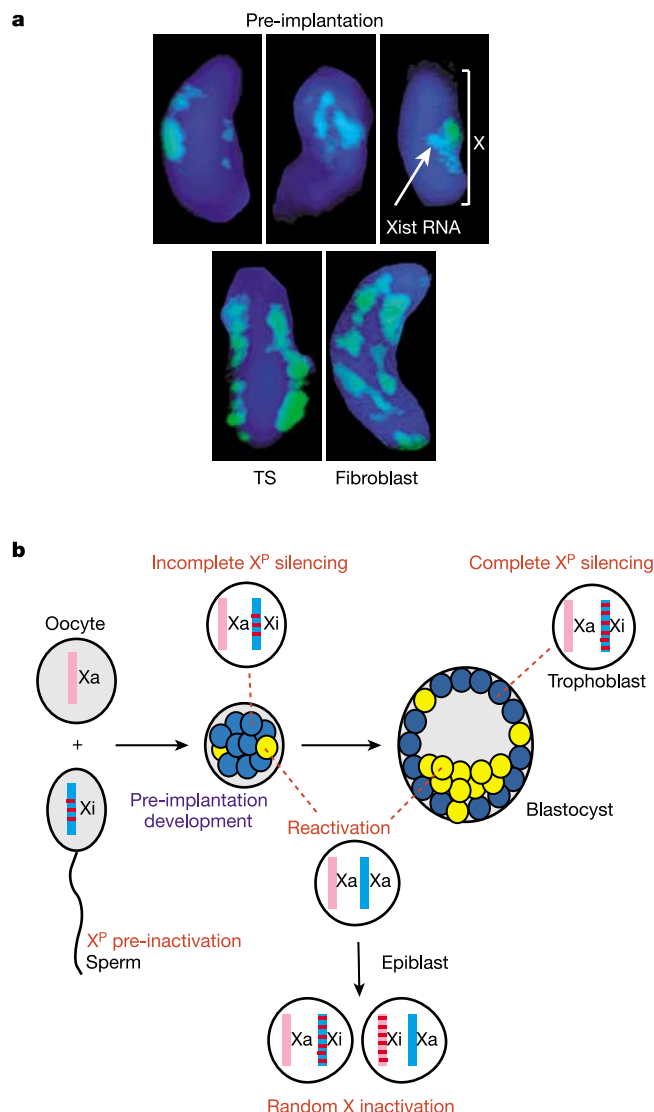


Figure 4 Limited spread of *Xist* RNA and a unified model of XCI in the mouse. **a**, *Xist* RNA-FISH of 8- and 16-cell embryos, TS cells, and fibroblasts. Shown is *Xist* RNA (green splotches) along the Xi chromosome (oblong blue body) in mitosis of representative blastomeres. **b**, A unified model. The female zygote inherits a pre-inactivated X^P (light blue) and an active X^M (pink) and maintains a partially silenced X^P (partial *Xist* coat) until trophoblast differentiation, whereupon *Xist* coating (red bars along the X chromosome) and silencing become extensive. The epiblast lineage and a small percentage of cells in the pre-implantation embryo (yellow circles) either spontaneously lose the imprint or purposefully undergo reactivation of the X^P (downregulation of *Xist*) to enable random XCI. Xa, active X chromosome.

expressed equally from cells known to carry active X^M and X^P and that, at the same time, they are subject to complete silencing in cells known to have imprinted XCI. These findings further highlight the unusual manifestation of XCI in pre-implantation embryos.

Thus, imprinted XCI seems to be a bi-phasic process, with leaky silencing occurring in the pre-implantation period, followed by more complete silencing in post-implantation cells. Previous studies have uncovered many changes in the Xi chromatin around the implantation period, including a shift from early to late replication timing²¹ and increased recruitment of heterochromatic factors such as Eed and Ezh2 (ref. 22). These changes may remodel the Xi for more complete silencing on a per-gene and a pan-chromosomal basis, consistent with our observation of a DAPI-intense Xi body in TS but not pre-implantation cells (data not shown). Finally, we asked whether the degree of silencing correlated with the spread of Xist RNA along the X chromosome. Analysis of metaphase chromosomes indicated that, although TS cells and somatic cells showed extensive Xist RNA banding^{10,18,23}, 8- and 16-cell embryos exhibited more limited spread of Xist RNA, with the RNA concentrated in the central third of the X chromosome (Fig. 4a). Thus, the extent of silencing correlated with the spread of Xist RNA on metaphase chromosomes.

Our work demonstrates that the XX embryo is largely dosage compensated from the time of conception, a conclusion that can be reconciled with the classical two Xa (active X chromosome) model. As with previous studies, we observed bi-allelic *Hprt* and *Gla* expression^{3–5}, but a critical difference is our observed bias against X^P . Like our study, the older studies found embryo-to-embryo variability, as evidenced by significant overlap in presumptive male and female activities^{3–5}. Furthermore, synchronous replication of the X chromosomes² does not conflict with our conclusion, because X-chromosome silencing can take place in ES cells without late-replication timing during a period of 'reversible' XCI²¹. We suggest that the pre-implantation XCI is also reversible, a necessity for later imprint erasure in the epiblast.

Several mechanisms can be proposed for pre-implantation XCI. In one, silencing is driven by the zygote at the 2-cell stage. The coincident upregulation of Xist RNA suggests an Xist-dependent mechanism. Alternatively, in light of meiotic sex chromosome inactivation during male meiosis²⁴, the X^P may actually be silenced in the paternal germ line and then transmitted to the zygote in a 'pre-inactivated' state¹¹ (Fig. 4b). Indeed, paternally driven XCI is a more parsimonious solution, as there seems to be a continuum of X^P silence from the paternal germ line to the zygote. Pre-inactivation allows fathers to ensure dosage equivalence of XX and XY progeny from the time of conception. It is also possible that father and zygote both have active roles; for example, the paternal germ line may initiate silencing (in an Xist-independent way⁸) and the zygote may maintain it during a time when the Xi state is not yet fixed. Pre-inactivation simplifies the ontogeny of XCI by eliminating multiple rounds of inactivation (germ line), reactivation (zygote), and re-inactivation (placenta, embryo proper) as postulated by the classical model.

The work argues that females are dosage compensated from conception to adulthood, with brief exceptions only during oogenesis and epiblast formation when both X chromosomes are active. The essential nature of XCI may explain why XX ES cells are karyotypically unstable. In light of this work, the unresolved question of whether imprinted XCI takes place in humans rises in importance. Imprinted XCI is thought to have evolved in marsupials^{6,25}. Because only XX offspring inherit an X^P , a simple solution for dosage compensation would have been to inactivate all X^P , a notion leading us to hypothesize that mammalian imprinting first arose on the X chromosome²⁶. Early on, silencing may have been restricted to genes lacking Y chromosome homologues near the *Xi*^{6,25,27}, with subsequent piecemeal spread along the X chromosome. Indeed, this proposed evolutionary sequence is mirrored in

early mouse development. As with pre-implantation mouse XCI, marsupial XCI also shows gene-to-gene and cell-to-cell variability^{25,27}. If ontogeny recapitulates phylogeny, the early mouse embryo would provide an excellent retrospective for the evolution of imprinting. □

Methods

Pre-implantation embryos

Fertilized eggs were recovered at 0.5 days post coitus from the corresponding matings and cultured *in vitro* in M16 (Sigma) or KSOM (Specialty Media) to the appropriate stages. For FISH analyses embryos were derived from crosses between B6CBAF1 females and B6CBAF1 males. For allele-specific expression analysis embryos were derived from crosses between *M. m. castaneus* (CAST/Ei) and *M. m. musculus* (C57BL/6). Embryos were sexed by Xist expression.

Cell lines

We derived new XX TS lines from 3.5 days post coitus blastocysts collected from a C57BL/6 × CAST/Ei mating (where the female was C57BL/6) as described²⁸. TS3.4 was clonally derived from a single colony of one embryo, whereas TS7 was a mixture of all TS colonies of a second embryo. The 16.7 female ES line is derived from a 129 × (CAST × 129) cross²⁹. 129 is a strain of *M. musculus* origin.

RNA-FISH

Pre-implantation embryos were treated with Tyrodes solution to remove the zona pellucida and allowed to recover in M2 medium (Sigma). Embryos were rinsed in PBS, and put through the following series of cold CSK buffer (100 mM NaCl, 300 mM sucrose, 10 mM PIPES pH 6.8, 3 mM MgCl₂): CSK for 60 s, CSK plus 0.5% Triton-X100 for 2–5 min, CSK for 60 s, followed by fixation in 4% paraformaldehyde for 10 min. Embryos were then dropped onto a glass slide, air-dried and stored in 70% EtOH. TS cells were cytopun onto a glass slide, and treated as described above for embryos. For metaphase chromosome analysis, embryos (stripped of the zona pellucida) and TS cells were swelled in 75 mM KCl followed by fixation in 4% paraformaldehyde plus 0.15% Triton-X100 for 10 min. RNA-FISH was performed essentially as described¹⁵. An Xist riboprobe cocktail was generated with Alexa Fluor 488-5-UTP (Molecular Probes). Cot-1 DNA probe was generated using the Prime-It Fluor Fluorescence Labeling Kit (Stratagene) with 1 μg mouse Cot-1 genomic DNA and Texas red-12-dUTP (Molecular Probes). Fluorescein isothiocyanate-labelled X chromosome painting probes were from Cambio. Fluorescence images were analysed as follows: z-series were acquired using Openlab software (Improvision) and three-dimensional deconvolution was performed with Velocity software (Improvision). Counting of Xist coating and Cot-1 holes for all blastomeres in an embryo was done by scanning through z-sections.

RT-PCR

Allele-specific RT-PCR is based on relative amplification differences between C57BL/6 and CAST/Ei amplicons with internal restriction-site SNPs, performed as described²⁹ except that RT-PCR was done using the One-Step RT-PCR kit (Invitrogen) for 40 cycles (94 °C, 48–58 °C, 72 °C). *Chic1* and *Fmr1* were amplified for an additional 20 cycles with nested primers. RT-PCR primers for all genes, with the exception of *Chic1*, spanned at least one intron, ensuring that PCR products were derived from RNA not genomic DNA. Minus RT reactions were included for *Chic1* to verify amplification of RNA and not DNA. One morula contained enough RNA to test up to four genes. Quantitative measures were obtained either by phosphorimaging (Molecular Dynamics) or from scanned autoradiographic images using NIH Image version 1.63 software. The following are forward and reverse RT-PCR primers (5' to 3'), Southern probes and polymorphic restriction sites (respectively) for the genes analysed. *Ptcl1* (ref. 30): CGACTTGATAGCGATGGAGC, CAATGGTTGGGTTGACAGGC, CCGTGTGGTGGATACCGAGTTC, *PvuII*; *Fmr1*: GAACAGCTGATGGATCCCTGC, AGTAAAACTCTG CCTGAAGTC (nested reverse primer GTGCAATCTGGAAGATTGTG), GACAAATAGTAGGCAAGATGGC, *HaeIII*; *Mecp2*: CATGGTAGCTGGGATGTTAGG, GCAATCAATTCTACTTTAGAGCG, GAAGGCAAGCATGAGCCACTACAACCTTCAG CCCACCAT, *Ddel*; *G6pd*: GGAGTGATGAACCTAGGGAAGC, ATGTAGCTGGGTTT ACTGGTGG, CGAGAAAAGCCCCAGCCATCC, *SerFI*; *Abc7*: AAGCATTGGCGAGTT CTGACC, TCTAGTATCAACATCCTTTAACC, GTAGCATCTAGTAACGTCTGGAG, *NlaIII*; *Chic1*: GAGTGCCCTTCTAATAAGTGG, CTGGAACTACTAGTAGCAGG (nested reverse primer TGAACCTCAAGAGATCTGTGGC), CTTCCTCTTCTGAAG TACACAG, *MseI*; *Xnp*: GAAGAAAAGGGATACCCCATGCG(G*)GCC, CCAGTTGGTATGTTGAACGC, GATACCATACTTGCAAGACTTC, *EagI*; *Atp7a*: GCGCTTCTATCTCTCTGTAG, GCAGCACATTAGCAACTTCTAAC, TGACACAGC ACTTACGTTGATG, *HpyCH4III*; *Pgk1*: CGTGATGAGGGTGGACTTCAAC, TAGTTTGGACAGTGAGGCTCGG, TGAACCTCAATCTCTGCTGGGC, *MseI*; *Rlim*: GAGCCCGAGTGAAATAGAGC, GGTGGCAGTCTCTGTACTGC, GACTGGAATACAAACAAGAGTGG, *HaeIII*; *Gla*: CCGTCCCCTACTCATGTCC(C*) A(G*)GATCT, GTCAAAGTCCATTATAGAAGCC, GGGCTGTGGCTGTGAGAAACC TGC, *AluNI*; *Pdhal*: GACCCTATTATGCTTCTCAAG, CACGCACTTCAAAGGGAGG ATAC(C*)GCT, AGAAATCGAGGATGCTGCCAG, *PvuII*; *Nxf1*: GGCGGACGAAGGAAGTCATAC, TGCTACATCGCCATCATCTTCC, GGGAATCGTA GGCTCGGAAGAG, *HpaII*; *Yyl*: GGTTCATCTGGAGAAAAGGCC, GGGTCGAGAAG GTCTTCT(G*)TCTTC, GGTTCATCTGGAGAAAAGGCC, *EarI*; *B2m*: TGACCGGCC TGTAGCTATCC, GATGCTTGATCATGTCTCG, GAGAATGGGAAGCCGAACATAC, *BglI*; *Hras*: AGACACAGCAGGTCAAGAAGAG, CATCGGGTGGGTTCAGTTTCC,

TCGAGGACATCCATCAGTACAG, *HaeII*; *Wars*: TTCCTCACTGACACGGCCAAG, CAGAAGCACTGGAAGTGAAGG, ATCAAGAGCAAGGTCAACAAAGC, *BstEII*. Designated primers for *Xnp*, *Pdha1*, *Gla* and *Yy1* contain engineered base changes (indicated by an asterisk) to facilitate allelic analysis. Changes do not affect relative priming efficiencies between the alleles, as the changes occur on both (details available on request). *Xist* and *Hprt* assays have been described²⁹.

Received 28 October; accepted 20 November 2003; doi:10.1038/nature02222.
Published online 7 December 2003.

1. Lyon, M. F. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* **190**, 372–373 (1961).
2. Mukherjee, A. B. Cell cycle analysis and X-chromosome inactivation in the developing mouse. *Proc. Natl Acad. Sci. USA* **73**, 1608–1611 (1976).
3. Epstein, C., Smith, S., Travis, B. & Tucker, G. Both X-chromosomes function before visible X-chromosome inactivation in female mouse embryos. *Nature* **274**, 500–503 (1978).
4. Kratzer, P. G. & Gartler, S. M. HGPRT activity changes in preimplantation mouse embryos. *Nature* **274**, 503–504 (1978).
5. Adler, D. A., West, J. D. & Chapman, V. M. Expression of alpha-galactosidase in preimplantation mouse embryos. *Nature* **267**, 838–839 (1977).
6. Lyon, M. F. In *Results and Problems in Cell Differentiation* (ed. Ohlsson, R.) 73–90 (Springer, Heidelberg, 1999).
7. Takagi, N. & Sasaki, M. Preferential inactivation of the paternally derived X-chromosome in the extraembryonic membranes of the mouse. *Nature* **256**, 640–642 (1975).
8. Marahrens, Y., Panning, B., Dausman, J., Strauss, W. & Jaenisch, R. Xist-deficient mice are defective in dosage compensation but not spermatogenesis. *Genes Dev.* **11**, 156–166 (1997).
9. Kay, G. F. *et al.* Expression of Xist during mouse development suggests a role in the initiation of X-chromosome inactivation. *Cell* **72**, 171–182 (1993).
10. Matsui, J., Goto, Y. & Takagi, N. Control of Xist expression for imprinted and random X-chromosome inactivation in mice. *Hum. Mol. Genet.* **10**, 1393–1401 (2001).
11. Huynh, K. D. & Lee, J. T. Imprinted X inactivation in eutherians: a model of gametic execution and zygotic relaxation. *Curr. Opin. Cell Biol.* **13**, 690–697 (2001).
12. Krietsch, W. K. *et al.* The expression of X-linked phosphoglycerate kinase in the early mouse embryo. *Differentiation* **23**, 141–144 (1982).
13. Latham, K. E., Patel, B., Bautista, F. D. M. & Hawes, S. M. Effects of X-chromosome number and parental origin on X-linked gene expression in preimplantation mouse embryos. *Biol. Reprod.* **63**, 64–73 (2000).
14. Mayer, W., Niveleau, A., Walter, J., Fundele, R. & Haaf, T. Demethylation of the zygotic paternal genome. *Nature* **3**, 501–502 (2000).
15. Hall, L. L. *et al.* An ectopic human XIST gene can induce chromosome inactivation in postdifferentiation human HT-1080 cells. *Proc. Natl Acad. Sci. USA* **99**, 8677–8682 (2002).
16. Schultz, R. M. Regulation of zygotic gene activation in the mouse. *Bioessays* **15**, 531–538 (1993).
17. Hartshorn, C., Rice, J. E. & Wang, L. J. Differential pattern of Xist RNA accumulation in single blastomeres isolated from 8-cell stage mouse embryos following laser zona drilling. *Mol. Reprod. Dev.* **64**, 41–51 (2003).
18. Mak, W. *et al.* Mitotically stable association of polycomb group proteins Eed and Enx1 with the inactive X-chromosome in trophoblast stem cells. *Curr. Biol.* **12**, 1016–1020 (2002).
19. Nothias, J. Y., Majumder, S., Kaneko, K. J. & DePamphilis, M. L. Regulation of gene expression at the beginning of mammalian development. *J. Biol. Chem.* **270**, 22077–22080 (1995).
20. Heard, E., Clerc, P. & Avner, P. X-chromosome inactivation in mammals. *Annu. Rev. Genet.* **31**, 571–610 (1997).
21. Wutz, A. & Jaenisch, R. A shift from reversible to irreversible X inactivation is triggered during ES cell differentiation. *Mol. Cell* **5**, 695–705 (2000).
22. Chadwick, B. P. & Willard, H. F. SETting the stage: Eed-Enx1 leaves and epigenetic signature on the inactive X-chromosome. *Dev. Cell* **4**, 445–447 (2003).
23. Duthie, S. M. *et al.* Xist RNA exhibits a banded localization on the inactive X-chromosome and is excluded from autosomal material in cis. *Hum. Mol. Genet.* **8**, 195–204 (1999).
24. Lifschytz, E. & Lindsley, D. L. The role of X-chromosome inactivation during spermatogenesis. *Proc. Natl Acad. Sci. USA* **69**, 182–186 (1972).
25. Graves, J. A. M. Mammals that break the rules: Genetics of marsupials and monotremes. *Annu. Rev. Genet.* **30**, 233–260 (1996).
26. Lee, J. T. Molecular links between X-inactivation and autosomal imprinting: X-inactivation as a driving force for the evolution of imprinting. *Curr. Biol.* **13**, R242–R254 (2003).
27. Jegalian, K. & Page, D. C. A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X-inactivated. *Nature* **394**, 776–780 (1998).
28. Tanaka, S., Kunath, T., Hadjantonakis, A. K., Nagy, A. & Rossant, J. Promotion of trophoblast stem cell proliferation by FGF4. *Science* **282**, 2072–2075 (1998).
29. Stavropoulos, N., Lu, N. & Lee, J. T. A functional role for Tsix transcription in blocking Xist RNA accumulation but not in X-chromosome choice. *Proc. Natl Acad. Sci. USA* **98**, 10232–10237 (2001).
30. Carrel, L. *et al.* X inactivation analysis and DNA methylation studies of the ubiquitin activating enzyme E1 and PCTAIRE-1 genes in human and mouse. *Hum. Mol. Genet.* **5**, 391–401 (1996).

Acknowledgements We are grateful to S. Shinwa for instruction in manipulating pre-implantation embryos, and to S. Shibata and Y. Ogawa for sharing reagents. We thank C. L. Tsai, L. F. Zhang and B. K. Sun for critical reading of the manuscript, and all members of the laboratory for discussion. This work was funded by the National Institutes of Health, USA, the Pew Scholar Program, and the Howard Hughes Medical Institute.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.T.L. (lee@frodo.mgh.harvard.edu).

Evolutionary conservation of biogenesis of β -barrel membrane proteins

Stefan A. Paschen¹, Thomas Waizenegger¹, Tincuta Stan¹, Marc Preuss¹, Marek Cyrklaff², Kai Hell^{1*}, Doron Rapaport^{1*} & Walter Neupert¹

¹Adolf-Butenandt-Institut für Physiologische Chemie, Ludwig-Maximilians-Universität München, Butenandtstrasse 5, D-81377 München, Germany

²Max-Planck-Institut für Biochemie, Abteilung für Molekulare Strukturbiochemie, Am Klopferspitz 18a, D-82152 Martinsried, Germany

* These authors contributed equally to this work

The outer membranes of mitochondria and chloroplasts are distinguished by the presence of β -barrel membrane proteins^{1,2}. The outer membrane of Gram-negative bacteria also harbours β -barrel proteins³. In mitochondria these proteins fulfil a variety of functions such as transport of small molecules (porin/VDAC), translocation of proteins (Tom40) and regulation of mitochondrial morphology (Mdm10)^{4–7}. These proteins are encoded by the nucleus, synthesized in the cytosol, targeted to mitochondria as chaperone-bound species, recognized by the translocase of the outer membrane, and then inserted into the outer membrane where they assemble into functional oligomers^{8–11}. Whereas some knowledge has been accumulated on the pathways of insertion of proteins that span cellular membranes with α -helical segments, very little is known about how β -barrel proteins are integrated into lipid bilayers and assembled into oligomeric structures¹². Here we describe a protein complex that is essential for the topogenesis of mitochondrial outer membrane β -barrel proteins (TOB). We present evidence that important elements of the topogenesis of β -barrel membrane proteins have been conserved during the evolution of mitochondria from endosymbiotic bacterial ancestors¹³.

To find components involved in the import, folding and assembly of β -barrel proteins we used a proteomic approach. On analysis by mass spectrometry of purified outer membranes of mitochondria from *Neurospora crassa*, we detected a protein of 55 kDa, hereafter called Tob55 (Supplementary Fig. S1a). Homologues are present in the genomes of virtually all eukaryotes (Supplementary Fig. S1b). This protein displayed significant sequence similarity with the Omp85 protein, which is present in the outer membrane of Gram-negative bacteria (Supplementary Fig. S1c)¹⁴. Tob55, like Omp85, is predicted to contain β -strands that are present mainly in the carboxy-terminal part of the protein. For further analysis we investigated Tob55 in *Saccharomyces cerevisiae*. Subcellular fractionation experiments revealed that Tob55 is indeed located in mitochondria (Supplementary Fig. S2). It represents an integral component of the mitochondrial outer membrane, probably exposing an amino-terminal domain to the intermembrane space (Supplementary Fig. S3).

Cells harbouring the *TOB55* gene under the control of the *GAL10* promoter were first grown on galactose-containing medium, and then shifted to galactose-free medium. Their growth slowed down at about 40 h after the shift (Fig. 1a). This is in agreement with previous reports demonstrating the essential nature of this gene¹⁵. Tob55 was virtually absent at 43 h after the shift (Tob55↓) (Fig. 1b). Mitochondria from cells isolated 47 h after the shift (Tob55↓) had very low levels of the β -barrel outer membrane proteins porin, Tom40 and Mdm10 for which the β -barrel structure was predicted (refs 5, 6; and S.A.P., unpublished results). The levels of other outer membrane proteins (Tom70, Tom20 and OM45) as well as those mitochondrial proteins residing in the intermembrane space

(CCPO), inner membrane (Tim17) or the matrix (aconitase) were essentially unaffected (Fig. 1b).

Tob55 $\downarrow$ mitochondria were analysed for their ability to import pre-proteins *in vitro* (Fig. 1c). They exhibited strongly reduced import of the β -stranded outer membrane proteins porin, Tom40 and Tob55 itself. On the other hand, proteins anchored to the outer membrane by α -helical segments (such as Tom20) were imported at roughly wild-type levels, as was the import of proteins destined to the intermembrane space (CCHL), the inner membrane (pDLD-DHFR) and the matrix (pF β). This suggests that Tob55 has an essential and specific role in the import of β -barrel proteins.

To determine the native molecular mass of Tob55, mitochondria were lysed with digitonin and proteins were analysed by blue native

gel electrophoresis (BN-PAGE) (Fig. 2a, left panel) and sizing chromatography (Fig. 2a, right panel). A similar apparent molecular mass of 220–250 kDa was obtained with both methods. Thus, Tob55 is part of an oligomeric complex called the topogenesis of mitochondrial outer membrane β -barrel proteins (TOB) complex. It is distinct from the translocase of the outer membrane (TOM) complex as it migrates differently in BN-PAGE assays, and antibodies against Tob55 precipitate the TOB complex but not the TOM complex (Fig. 2a, left and middle panels).

We expressed a His-tagged version of Tob55 in *Escherichia coli*, purified it from inclusion bodies by Ni-NTA chromatography, solubilized the protein in 8 M urea and refolded it. The circular dichroism (CD) spectrum of the refolded Tob55 revealed a relatively high content of β -strands (approximately 28%), comparable with

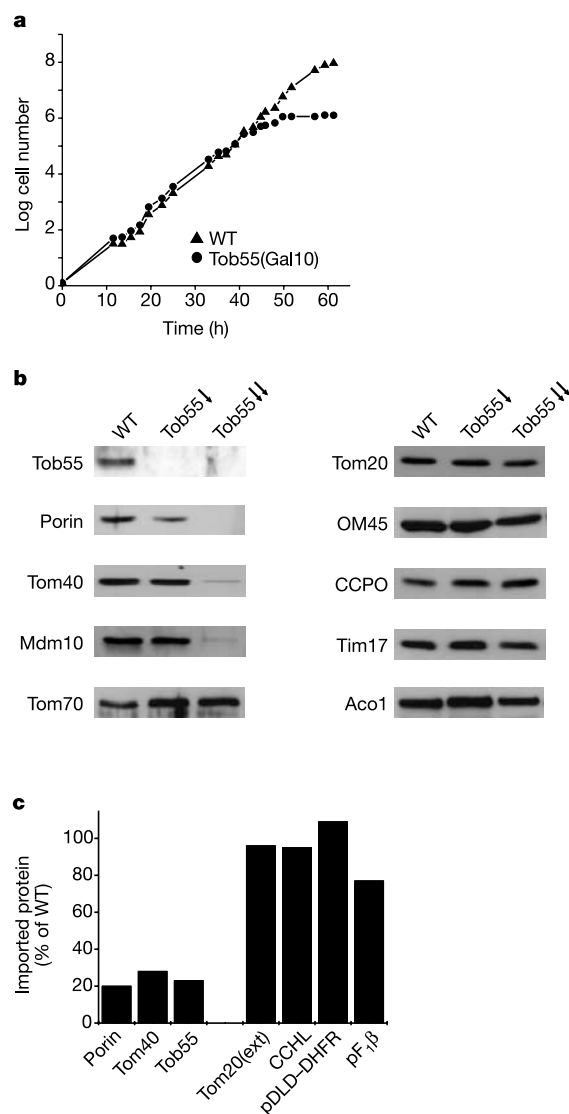


Figure 1 Tob55 is essential for the biogenesis of β -barrel proteins. **a**, Downregulation of Tob55 affects cell growth. Cells from a wild-type strain (WT) or from a strain expressing Tob55 under the control of the *GAL10* promoter (Tob55(Gal10)) were shifted from galactose- to glucose-containing medium at time zero. **b**, Levels of β -barrel proteins are reduced in mitochondria from cells depleted of Tob55 for 43 h (Tob55 $\downarrow$) or 47 h (Tob55 $\downarrow\downarrow$), as determined by immunodecoration. **c**, Tob55 is required for the import of β -barrel proteins. Radiolabelled pre-proteins were imported into wild-type or Tob55 $\downarrow$ mitochondria. For each precursor, the results of at least three experiments are within $\pm 15\%$ of the indicated average value. Tom20(ext), Tom20 variant with N-terminal extension, adopting wild-type topology.

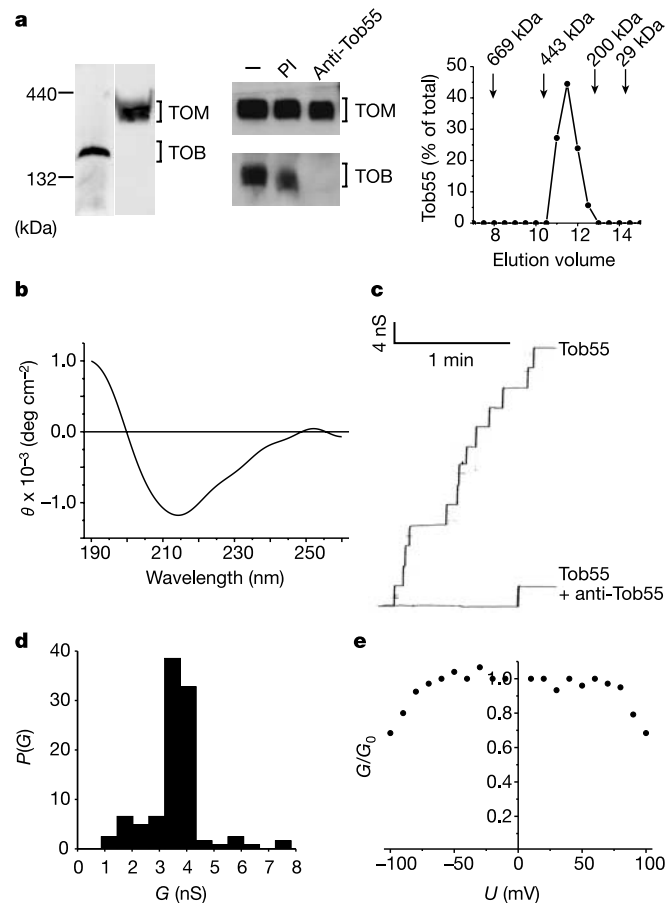


Figure 2 Tob55 contains β -stranded segments, is present as an oligomeric assembly, and forms channels in lipid bilayers. **a**, Tob55 occurs in a complex of approximately 220–250 kDa that is distinct from the TOM complex. Left panel: mitochondria analysis by BN-PAGE and immunodecoration with antibodies against Tob55 (left lane) or Tom40 (right lane). Middle panel: mitochondria were lysed with digitonin and added to protein A Sepharose beads with pre-bound Tob55 antibodies or pre-immune serum (PI). Analysis was carried out as above. Right panel: gel filtration and immunoblotting of fractions with antibodies against Tob55. **b**, Far-ultraviolet CD spectrum of refolded recombinant Tob55 (evaluation of the spectrum suggests 28% β -sheet, 24% α -helix and 12.5% turn structure). θ , ellipticity. **c**, Single-channel recordings of Tob55. Recombinant Tob55 was added directly or after pre-incubation with anti-Tob55 IgG to a black lipid membrane. **d**, Histogram of channel conductances. P(G) is the probability that a given conductance increment G was observed ($n = 122$ analysed). **e**, Voltage dependence of channel conductances. The ratio of the conductance, G, at a given voltage, U, divided by G $_0$ at 10 mV is shown as a function of voltage.

that found in Tom40 and porin⁶ (Fig. 2b). The refolded Tob55 inserted spontaneously into lipid bilayers (Fig. 2c). It formed channels with an average conductivity of 3.7 nS at 1 M KCl (Fig. 2d). The channels partially closed at applied voltages higher than ± 70 mV (Fig. 2e). Thus, the Tob55 channel differs in its conductivity and its voltage-dependence from mitochondrial and bacterial porins^{3,4,6}.

We also isolated the His-tagged version of Tob55 by Ni-NTA chromatography after lysis of mitochondria from the yeast strain Tob55(Gal10) with *n*-dodecyl- β -D-maltoside (DDM). Both native and recombinant Tob55 were analysed by electron microscopy after cryo-negative staining¹⁶ (Fig. 3a–d). Ring-shaped assemblies were predominant. Most of the rings enclosed a smaller density located either centrally or slightly off-centre. Preliminary two-dimensional average maps (Fig. 3e, f) revealed an outer diameter of approximately 15 nm and an inner diameter of about 7–8 nm. The central density measured approximately 4–5 nm. The rings exhibited a five-fold rotational symmetry (Fig. 3g, h).

In order to identify the molecular function of the TOB complex, mitochondria carrying His-tagged Tob55 were incubated with radiolabelled precursors of β -barrel proteins at low temperature to accumulate intermediates^{17–19}. During the import the precursors became associated with Tob55 (Fig. 4a). In contrast to the precursor of porin in transit, endogenous fully assembled porin as well as the precursor of Tom20—an outer membrane protein with a single α -helical transmembrane segment—were not co-purified with Tob55 (Fig. 4a). This shows that Tob55 interacts with precursors of β -barrel proteins on their import pathways.

To determine at which stage Tob55 is involved in the import

process, mitochondria after the import of various β -barrel proteins were subjected to BN-PAGE, which leads to separation of import intermediates. In the case of the Tom40 precursor, 250 kDa and 100 kDa complexes containing intermediates were reported that kinetically precede formation of the mature TOM complex^{17,19}. These intermediates were absent after import into mitochondria isolated from Tob55-depleted cells (Fig. 4b). We chose conditions that favour accumulation of the 250 kDa and 100 kDa intermediate complexes, and investigated whether they contained Tob55. To this end, mitochondria with accumulated intermediates were lysed, incubated with antibodies against Tob55 and then analysed by BN-PAGE. The 250 kDa but not the 100 kDa intermediate was shifted to higher apparent molecular mass by formation of a supercomplex with the antibodies (Fig. 4c, left panel). When the antibodies were pre-incubated with an excess of Tob55 antigen, the shift was efficiently inhibited (Fig. 4c, middle panel). The presence of Tob55 in the 250 kDa intermediate was also demonstrated by the ability of antibodies against Tob55 to deplete this species from a lysate of mitochondria (Fig. 4c, right panel). A similar behaviour was observed with the precursor of Mdm10 (Fig. 4d). Thus, Tob55 is an integral component of import intermediates of β -barrel proteins. Mdm10 and Tom40 precursors associated with Tob55 were largely resistant to degradation upon treatment of intact mitochondria with proteases (Fig. 4e and data not shown), in agreement with previous observations on the 250 kDa intermediate of Tom40 (ref. 19). This suggests that the Tob55-associated precursor had largely or completely crossed the outer membrane.

To analyse whether this translocation occurs via the TOM complex, we blocked the latter by addition of an excess of recombinant pSu9(1–69)-DHFR, a precursor consisting of the targeting signal of ATP synthase subunit 9 and dihydrofolate reductase¹⁸. Then porin was imported and Tob55 was isolated. Formation of the Tob55-containing intermediate was strongly reduced when the TOM complexes were occupied (Fig. 4f). A similar effect was observed when Mdm10 was imported (Fig. 4g). Thus, β -barrel proteins use the TOM complex before they interact with Tob55. In fact, the TOB complex seems to be essential for complete translocation across the TOM complex. In Tob55-depleted mitochondria the 250 kDa intermediate was not formed (Fig. 4b) and Tom40 precursor did not reach a protease-protected state (Fig. 4h).

A recent study suggested the participation of Mas37 in the sorting and assembly of outer membrane proteins in yeast²⁰. As with Tob55, Mas37 was found to be present in the 250 kDa assembly intermediate in the Tom40 import pathway. The TOB complex in Mas37-deleted cells was reduced in size by approximately 40 kDa (Supplementary Fig. S4a). Thus, Mas37 probably occurs in a complex with Tob55, and a Tob55-containing core complex is formed in its absence. This TOB core complex maintained the ability to form the assembly intermediate of Tom40 (Supplementary Fig. S4b). Furthermore, the TOB core complex is able to act as a functional insertion and assembly machinery, as cells deficient in Mas37 are viable. Notably, no homologues of Mas37 were found in bacteria or eukaryotes other than yeast. Mas37 was also reported to interact with pre-protein receptors of the outer membrane²¹. Thus, it might have a stabilizing effect on several outer membrane proteins. Extensive studies will be needed to understand its precise role.

Why would β -barrel proteins of the mitochondrial membrane undergo such a complicated topogenic pathway? It seems reasonable to assume that β -barrel proteins of mitochondria are evolutionarily derived from bacterial outer membrane β -barrel proteins^{13,22}. In view of the complex requirements for folding and assembly, it seems possible that the mitochondrial β -barrel proteins must follow a pathway conserved during evolution. In bacteria, β -barrel proteins insert from the periplasmic side into the outer membrane³. Recently, a major role in this process in *Neisseria meningitidis* was attributed to Omp85 (ref. 14). Accumulation of

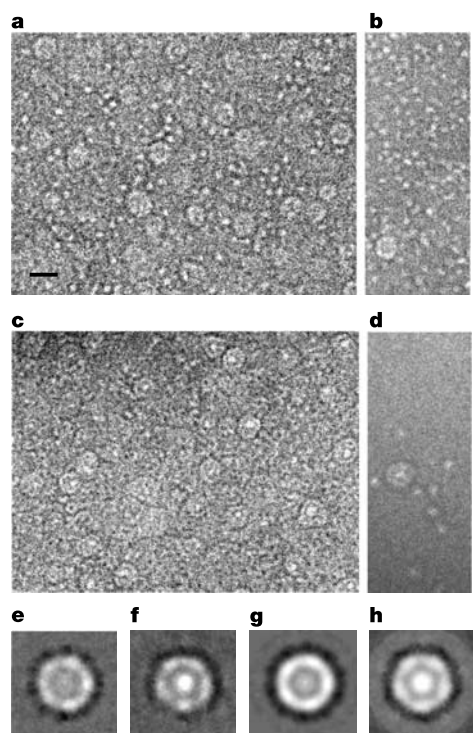


Figure 3 Electron microscopy of native and recombinant refolded TOB complexes using cryo-negative staining with ammonium molybdate. **a–d**, Unprocessed fields. **a, b**, Images of native TOB complexes attached to carbon support (**a**) and present over a hole in an EM-carbon grid (**b**). **c, d**, Same as **a** and **b**, but for recombinant refolded Tob55. **e–h**, Gallery of two-dimensionally averaged images; 15-nm particles appear most frequently in both preparations. **e**, Average of 347 particles of native Tob55. **f**, Average of 207 particles of the recombinant refolded Tob55. **g, h**, Same as **e** and **f**, but after imposing five-fold symmetry. Scale bar in **a** is 20 nm and pertains to **a–d**. The edges in **e–h** are 27 nm.

β -barrel intermediates in the periplasmic space of Omp85-deficient cells was presented as the evidence for this hypothesis¹⁴, although a contradictory report postulated a role of Omp85 in lipid export²³. Sequence similarity of Tob55 to the apparently ubiquitous Omp85-like proteins of Gram-negative bacteria¹⁴ is particularly high with respect to *Rickettsia prowazekii* (Supplementary Fig. S1c). The latter belongs to the α -Proteobacteria, which are thought to be evolutionary ancestors of mitochondria¹³. Not only structure but also topology and function of the bacterial Omp85 machinery may be conserved in the TOB complex. This would require that β -barrel precursors are also presented in mitochondria from the inner face of the outer membrane. Thus, the precursors first have to cross the outer membrane via the TOM complex, in order to then follow a conserved pathway of translocation into the outer membrane with

the help of the TOB complex.

As the TOB complex forms a channel with an apparently large diameter it is tempting to speculate that the channel may be large enough to accommodate up to 16–22 β -strands, usually present in β -barrel proteins³. The β -strands could fold into their native or near-native conformation in the channel, then to be released by lateral opening into the lipid phase of the outer membrane. Thus, the TOB complex might be viewed as an Anfinsen-type cage, in which the β -stranded membrane proteins can fold²⁴.

The existence of a channel of large size in the mitochondrial outer membrane has been suggested on the basis of electrophysiological experiments²⁵. The TOB complex might indeed be the complex of quest. Given that the Tob55 channel is large enough to allow the passage of smaller folded proteins, it would be interesting to study

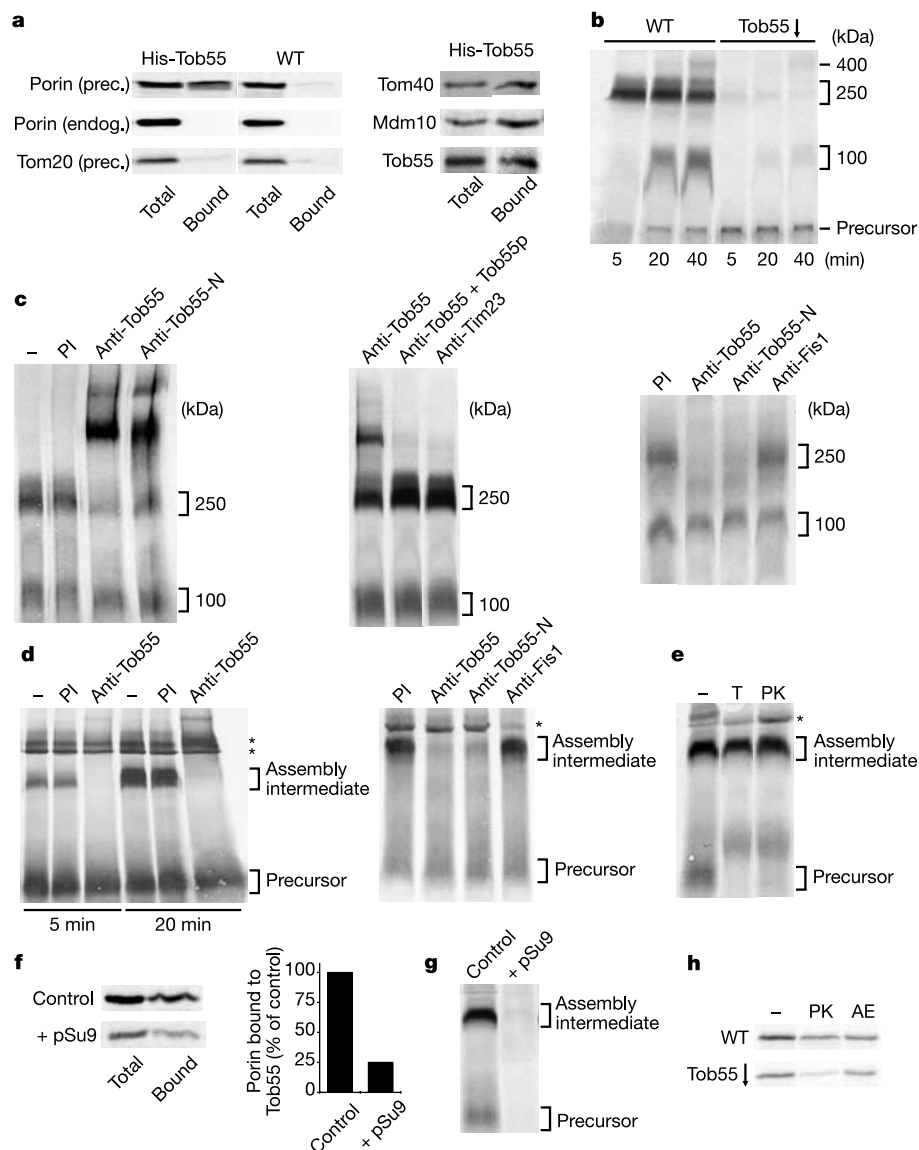


Figure 4 The TOB complex engages β -barrel proteins after their interaction with the TOM complex. **a**, Tob55 interacts with precursors (prec.) of β -barrel proteins. Radiolabelled pre-proteins were imported into the indicated mitochondria followed by Ni-NTA affinity purification. Total, 15% of input. Authentic porin (endog.) was immunodecorated. **b**, Tob55 is required for the import and assembly of Tom40. Import of radiolabelled Tom40 was analysed by BN-PAGE. **c**, Tob55 is part of the 250 kDa assembly intermediate of imported Tom40. Left and middle panels: lysed mitochondria were incubated with antibodies or with antigen-saturated antibodies (anti-Tob55 + Tob55p). Right panel: mitochondrial lysate was added to Sepharose beads containing the indicated

antibodies. PI, pre-immune serum. **d**, Tob55 is present in an assembly intermediate of Mdm10. Left panel: antibody shift; right panel: antibody depletion. Asterisks indicate unspecific bands. **e**, The assembly intermediate of Mdm10 is protected from proteases. T, trypsin; PK, proteinase K. **f**, **g**, β -Barrel precursors (**f**, porin precursor as in **a**; **g**, Mdm10) are transported to the TOB complex via the TOM complex. De-energized mitochondria were pre-incubated without (control) or with excess amounts of pSu9(1–69)–DHFR (pSu9). **h**, Tom40 precursor is protease-sensitive in Tob55Δ mitochondria. AE, alkaline extraction.

its possible additional role as a regulated conduit for the release of cytochrome *c*, a key step in apoptosis. □

Methods

Yeast strains

The haploid Tob55(Gal10) strain, expressing a Tob55 protein with an N-terminal octahistidine tag under control of the *GAL10* promoter, was constructed by transformation of the wild-type strain YPH499 with a polymerase chain reaction fragment and homologous recombination²⁶. The growth rate of this strain was similar to that of the wild-type strain. For depletion of Tob55, Tob55(Gal10) cells were shifted from lactate medium containing 0.1% galactose to lactate medium with 0.1% glucose.

Expression of recombinant Tob55 and refolding

A His-tagged version of Tob55 was cloned into the expression vector pQE60 (Qiagen) and expressed in *E. coli*. The protein was purified from inclusion bodies by Ni-NTA chromatography and eluted with 8 M urea, 50 mM Na-phosphate, 200 mM imidazole and 1 mM PMSF, pH 8.0. The purified protein was refolded by dialysis against decreasing concentrations of urea in 50 mM Na-phosphate, 50 mM NaCl and 0.05% DDM, pH 8.0.

CD spectroscopy

Far-ultraviolet CD measurements were performed using a Jasco J-715 spectrometer. Final CD spectra were obtained by averaging eight consecutive scans, and the secondary structure content was estimated²⁷.

Purification of native Tob55

Mitochondria were solubilized in 50 mM Na-phosphate, 300 mM NaCl, 20 mM imidazole, 10% glycerol, 1 mM PMSF and 1% DDM, pH 8.0, for 30 min on ice. After a clarifying spin, solubilized material was added to Ni-NTA agarose beads. After 1 h agitation at 4 °C, beads were washed with the same buffer containing 0.1% DDM and bound protein was eluted with 300 mM imidazole. For co-purification of *in vitro* imported precursor proteins with Tob55, mitochondria (250 µg) were treated as described above with the exception that digitonin was used as detergent.

Electron microscopy

Samples for electron microscopy were prepared using cryo-negative staining in ammonium molybdate¹⁶. The electron microscopy grids were imaged on the CM200-FEG electron microscope. Defocus pairs of 0.8 and 1.6 µm were recorded at the nominal magnification of 38,000. The images were further analysed using the EM software package²⁸. Particles for further analysis were interactively selected and two-dimensional average maps were calculated without correction for contrast transfer function. A first reference was created by reference-free alignment and then subjected to iterative rotational and translational alignment. The particles with lowest correlation coefficient to the current reference were gradually rejected. Resulting average maps were subjected to rotational analysis by imposing the rotational symmetries in a range between two- to ten-fold. The best symmetry fit was determined by eye.

Antibody-shift/depletion of protein complexes

Radiolabelled precursor proteins were incubated with mitochondria. Samples were re-isolated and solubilized in buffer containing 1% digitonin. In the case of the antibody shift experiments, immunoglobulin-γ (IgG) was added directly to the mitochondrial lysate and incubated for 1 h at 4 °C. For the immunodepletion experiments mitochondrial lysate was added to protein A Sepharose beads that were pre-incubated with the indicated antibodies. After 1 h incubation at 4 °C the beads were removed. The supernatants of the antibody depletion and the samples of the antibody shift were analysed by BN-PAGE¹⁷.

Miscellaneous

In vitro import of radiolabelled pre-proteins into isolated yeast mitochondria was performed essentially as described^{17,18}. The 'black lipid bilayer' experiments were performed according to published procedures²⁹. Recombinant pSu9(1–69)-DHFR was purified as described¹⁸. Antibodies were raised in rabbits against the purified recombinant Tob55 protein (anti-Tob55) and a peptide covering amino acid residues 1–15 (anti-Tob55-N).

Received 17 August; accepted 30 October 2003; doi:10.1038/nature02208.

- Rapaport, D. Biogenesis of the mitochondrial TOM complex. *Trends Biochem. Sci.* **27**, 191–197 (2002).
- Schleiff, E. *et al.* Prediction of the plant β-barrel proteome: A case study of the chloroplast outer envelope. *Protein Sci.* **12**, 748–759 (2003).
- Schulz, G. E. The structure of bacterial outer membrane proteins. *Biochim. Biophys. Acta* **1565**, 308–317 (2002).
- Blachly-Dyson, E. & Forte, M. VDAC channels. *IUBMB Life* **52**, 113–118 (2001).
- Hill, K. *et al.* Tom40 forms the hydrophilic channel of the mitochondrial import pore for preproteins. *Nature* **395**, 516–521 (1998).
- Ahting, U. *et al.* Tom40, the pore-forming component of the protein-conducting TOM channel in the outer membrane of mitochondria. *J. Cell Biol.* **153**, 1151–1160 (2001).
- Sogo, L. F. & Yaffe, M. P. Regulation of mitochondrial morphology and inheritance by Mdm10p, a protein of the mitochondrial outer membrane. *J. Cell Biol.* **126**, 1361–1373 (1994).
- Endo, T. & Kohda, D. Functions of outer membrane receptors in mitochondrial protein import. *Biochim. Biophys. Acta* **1592**, 3–14 (2002).
- Matouschek, A. & Glick, B. S. Barreling through the outer membrane. *Nature Struct. Biol.* **8**, 284–286 (2001).
- Pfanner, N. & Geissler, A. Versatility of the mitochondrial protein import machinery. *Nature Rev. Mol. Cell Biol.* **2**, 339–349 (2001).
- Paschen, S. A. & Neupert, W. Protein import into mitochondria. *IUBMB Life* **52**, 101–112 (2001).
- Gabriel, K., Buchanan, S. K. & Lithgow, T. The alpha and the beta: protein translocation across mitochondrial and plastid outer membranes. *Trends Biochem. Sci.* **26**, 36–40 (2001).
- Gray, M. W., Burger, G. & Lang, B. F. Mitochondrial evolution. *Science* **283**, 1476–1481 (1999).
- Voulhoux, R., Bos, M. P., Geurtsen, J., Mols, M. & Tommassen, J. Role of a highly conserved bacterial protein in outer membrane protein assembly. *Science* **299**, 262–265 (2003).
- Brachat, A. *et al.* Analysis of deletion phenotypes and GFP fusions of 21 novel *Saccharomyces cerevisiae* open reading frames. *Yeast* **16**, 241–253 (2000).
- De Carlo, S., El-Bez, C., Alvarez-Rua, C., Borge, J. & Dubochet, J. Cryo-negative staining reduces electron-beam sensitivity of vitrified biological particles. *J. Struct. Biol.* **138**, 216–226 (2002).
- Rapaport, D. & Neupert, W. Biogenesis of Tom40, core component of the TOM complex of mitochondria. *J. Cell Biol.* **146**, 321–331 (1999).
- Krimmer, T. *et al.* Biogenesis of porin of the outer mitochondrial membrane involves an import pathway via receptors and the general import pore of the TOM complex. *J. Cell Biol.* **152**, 289–300 (2001).
- Model, K. *et al.* Multistep assembly of the protein import channel of the mitochondrial outer membrane. *Nature Struct. Biol.* **8**, 361–370 (2001).
- Wiedemann, N. *et al.* Machinery for protein sorting and assembly in the mitochondrial outer membrane. *Nature* **424**, 565–571 (2003).
- Gratzer, S. *et al.* Mas37p, a novel receptor subunit for protein import into mitochondria. *J. Cell Biol.* **129**, 25–34 (1995).
- Marcotte, E. M., Xenarios, I., van Der Blik, A. M. & Eisenberg, D. Localizing proteins in the cell from their phylogenetic profiles. *Proc. Natl Acad. Sci. USA* **97**, 12115–12120 (2000).
- Genevris, S., Steeghs, L., Roholl, P., Letesson, J. J. & van der Ley, P. The Omp85 protein of *Neisseria meningitidis* is required for lipid export to the outer membrane. *EMBO J.* **22**, 1780–1789 (2003).
- Ellis, R. J. Revisiting the Anfinsen cage. *Fold. Des.* **1**, R9–R15 (1996).
- Pavlov, E. V. *et al.* A novel, high conductance channel of mitochondria linked to apoptosis in mammalian cells and Bax expression in yeast. *J. Cell Biol.* **155**, 725–731 (2001).
- Lafontaine, D. & Tollervey, D. One-step PCR mediated strategy for the construction of conditionally expressed and epitope tagged yeast proteins. *Nucleic Acids Res.* **24**, 3469–3471 (1996).
- Raussen, V., Ruyschaert, J. M. & Goormaghtigh, E. Protein concentration is not an absolute prerequisite for the determination of secondary structure from circular dichroism spectra: a new scaling method. *Anal. Biochem.* **319**, 114–121 (2003).
- Hegerl, R. The EM Program Package: A platform for image processing in biological electron microscopy. *J. Struct. Biol.* **116**, 30–34 (1996).
- Popp, B., Court, D. A., Benz, R., Neupert, W. & Lill, R. The role of the N and C termini of recombinant *Neurospora* mitochondrial porin in channel formation and voltage-dependent gating. *J. Biol. Chem.* **271**, 13593–13599 (1996).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank U. Gärtner and P. Heckmeyer for technical assistance; L. Peters, T. Silberzahn and A. Weinzierl for their help in some experiments; S. Schmitt for outer membranes of *N. crassa* mitochondria; and C. Bornhövd for help in subcellular yeast fractionation. We are grateful to W. Baumeister for support and access to the electron microscope facility. We thank E. Weyher-Stingl for help with the CD measurements, and R. Benz and S. Nussberger for support with the electrophysiological experiments. This work was supported by the Deutsche Forschungsgemeinschaft (D.R.), Sonderforschungsbereich 594, the Bundesministerium für Bildung und Forschung (MITOP), the Fonds der Chemischen Industrie (W.N.), and predoctoral fellowships from the Boehringer Ingelheim Fonds (S.A.P. and T.W.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to W.N. (Neupert@bio.med.uni-muenchen.de).

An ABC transporter with a secondary-active multidrug translocator domain

Henrietta Venter, Richard A. Shilling, Saroj Velamakanni, Lekshmy Balakrishnan & Hendrik W. van Veen

Department of Pharmacology, University of Cambridge, Tennis Court Road, Cambridge CB2 1PD, UK

Multidrug resistance, by which cells become resistant to multiple unrelated pharmaceuticals, is due to the extrusion of drugs from the cell's interior by active transporters such as the human multidrug resistance P-glycoprotein¹. Two major classes of transporters mediate this extrusion^{2,3}. Primary-active transporters are dependent on ATP hydrolysis, whereas secondary-active trans-

porters are driven by electrochemical ion gradients that exist across the plasma membrane. The ATP-binding cassette (ABC) transporter LmrA⁴ is a primary drug transporter in *Lactococcus lactis* that can functionally substitute for P-glycoprotein in lung fibroblast cells⁵. Here we have engineered a truncated LmrA protein that lacks the ATP-binding domain. Surprisingly, this truncated protein mediates a proton–ethidium symport reaction without the requirement for ATP. In other words, it functions as a secondary-active multidrug uptake system. These findings suggest that the evolutionary precursor of LmrA was a secondary-active substrate translocator that acquired an ATP-binding domain to enable primary-active multidrug efflux in *L. lactis*.

LmrA and P-glycoprotein are both composed of two half-transporters, each containing an amino-terminal membrane domain (MD) and a carboxy-terminal nucleotide-binding domain (NBD). These half-transporters are covalently linked in P-glycoprotein and are present in a non-covalent homodimeric complex in LmrA⁶. To study the interaction between drugs and the MD of LmrA in the absence of the NBD, N-terminally His₆-tagged LmrA was truncated in the linker region that connects the MD to the NBD, downstream of intracellular domain 3 in accordance with the three-dimensional structure of the homologous MsbA monomer^{7,8}. Truncated LmrA (LmrA-MD) was detected as a 36-kDa protein by SDS–polyacrylamide-gel electrophoresis (SDS–PAGE) and immunoblots probed with anti-His₆-tag antibody (Fig. 1a). LmrA-MD shared structural properties with LmrA: in the presence of the crosslinker disuccinimidyl glutarate significant amounts of the LmrA-MD dimer were observed (Fig. 1a). In addition, complete proteolytic removal of the His₆ tag from LmrA-MD was achieved in inside-out vesicles, but not in right-side-out membrane vesicles, showing the intracellular location of the N terminus, as observed for LmrA⁶ (Fig. 1b).

The pharmacological properties of LmrA-MD were compared with those known for LmrA^{5,9}. Photoaffinity labelling of LmrA-MD with [³H]azidopine, an LmrA substrate, was inhibited by the LmrA functional inhibitor nicardipine, providing evidence that drug–protein interactions were retained in LmrA-MD (Fig. 2a). Surprisingly, in cytotoxicity assays LmrA-MD expression increased the sensitivity of *L. lactis* to LmrA substrates such as (1) cytotoxic agents including Hoechst 33342, ethidium, tetramethyl rosamine and

benzalkonium, (2) antibiotics including tetracycline and erythromycin, and (3) the anthracycline daunomycin (Fig. 2b, c). The sensitivity of LmrA-MD-expressing *L. lactis* cells to ethidium was completely reversed in the presence of the LmrA substrate vinblastine or the LmrA modulator verapamil (data not shown). When *L. lactis* was allowed to generate metabolic energy using glucose, cells expressing LmrA-MD exhibited an enhanced initial ethidium influx rate of 11.78 ± 0.28 arbitrary units (a.u.) min⁻¹; control cells exhibited a rate of 6.70 ± 0.15 a.u. min⁻¹ ($n = 8$) (Fig. 2d). The increased sensitivity of *L. lactis* cells towards multiple drugs was therefore the result of LmrA-MD-mediated drug uptake.

The mechanism of energy coupling to drug uptake by LmrA-MD in glucose-energized cells was investigated by using ethidium. The magnitude and composition of the protonmotive force (Δp) (interior alkaline and negative) in these cells was manipulated through the addition of the ionophores nigericin and valinomycin (Fig. 3a)¹⁰. Nigericin mediates the antiport of H⁺ and K⁺ down their concentration gradients, thereby selectively dissipating the transmembrane H⁺ gradient (ΔpH) in an electroneutral manner. Valinomycin mediates electrogenic uniport of K⁺, allowing the electrophoretic uptake of K⁺ in cells with dissipation of the transmembrane potential ($\Delta \psi$). Remarkably, a significant LmrA-MD-mediated ethidium uptake was observed in cells in the presence of the ΔpH and/or the $\Delta \psi$, but not in the absence of both (Fig. 3a), suggesting that the Δp is a driving force for ethidium transport by LmrA-MD.

To further study the role of the ΔpH and $\Delta \psi$ in LmrA-MD-mediated transport, ethidium uptake was measured in DNA-loaded proteoliposomes containing purified and functionally reconstituted LmrA-MD, in which the ΔpH and $\Delta \psi$ were imposed artificially. Studies on the accessibility of the His₆ tag in LmrA-MD in the proteoliposomes through protease cleavage showed that more than 96% of the protein molecules were oriented inside-out (data not shown). When potassium acetate-containing proteoliposomes were diluted 100-fold into a buffer containing potassium sulphate, a ΔpH

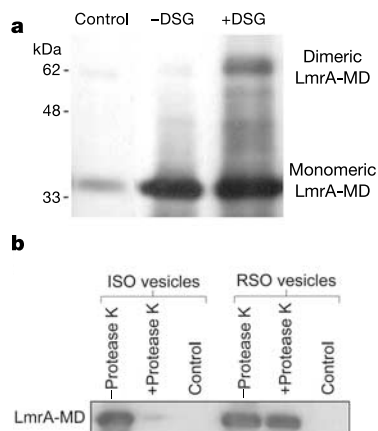


Figure 1 Structural properties of LmrA are maintained in LmrA-MD. **a**, Protein crosslinking with disuccinimidyl glutarate (DSG) in membrane vesicles without (Control) or with LmrA-MD, probed by immunoblotting with antibody against the His₆ tag. The migration of molecular mass markers is indicated. **b**, Immunodetection of residual N-terminal cytosolic His₆ tag in LmrA-MD before and after the digestion of inside-out (ISO) and right-side-out (RSO) membrane vesicles with protease K. To make the internal His₆ tag accessible, 1% Triton X-100 was added to the membrane vesicles before proteolysis (Control).

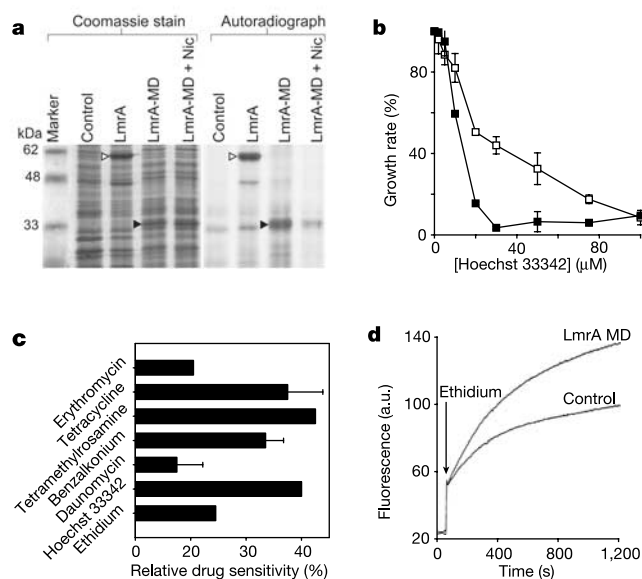


Figure 2 LmrA-MD mediates multidrug uptake in *L. lactis*. **a**, Photoaffinity labelling of LmrA (open arrowhead) and LmrA-MD (filled arrowhead) in inside-out vesicles with [³H]azidopine was detected on Coomassie-stained SDS–PAGE by autoradiography. Nic, 5 μM LmrA modulator nicardipine. **b**, Relative growth rate of cells expressing LmrA-MD (filled squares) and control cells (open squares) as a function of the Hoechst 33342 concentration ($P < 0.05$, $n = 4$). **c**, Drug sensitivity of LmrA-MD expressing cells relative to control cells (100%) ($P < 0.05$, $n = 4$). **d**, Ethidium uptake in cells expressing LmrA-MD or in control cells in the presence of 20 mM glucose.

(interior alkaline) was imposed across the lipid bilayer by the diffusion of acetic acid down its concentration gradient. In the presence of the imposed ΔpH , the formation of an inverted $\Delta\psi$ was prevented through the addition of valinomycin. Upon the 100-fold dilution of the proteoliposomes in a potassium-free acetate-containing buffer in the presence of valinomycin, a $\Delta\psi$ (interior negative) was imposed by electrogenic, downhill diffusion of K^+ .

In the presence of the imposed ΔpH , proteoliposomes containing inside-out-oriented LmrA-MD accumulated ethidium at an initial rate of $20.2 \pm 0.5 \text{ a.u. min}^{-1}$ versus $7.5 \pm 0.6 \text{ a.u. min}^{-1}$ ($n = 4$) in empty liposomes (control), and to a steady-state level that was 4.2 ± 0.3 -fold ($n = 4$) above the equilibration level in empty liposomes (Fig. 3b). In the presence of the imposed $\Delta\psi$, LmrA-MD-containing proteoliposomes accumulated ethidium 2.2 ± 0.1 -fold ($n = 4$) compared with empty liposomes, with an initial rate of $28.3 \pm 1.3 \text{ a.u. min}^{-1}$ versus $10.8 \pm 0.1 \text{ a.u. min}^{-1}$ ($n = 4$), respectively (Fig. 3c). The drug accumulation levels obtained were similar to those reported for ATP-dependent drug transport by LmrA⁴ and P-glycoprotein^{11,12}. Taken together, the results point to the LmrA-MD-mediated transport of ethidium in a Δp -dependent fashion.

To test whether variations in the amplitude or stability of the artificially imposed ΔpH and $\Delta\psi$ could account for the demonstrable differences in initial rates and steady-state levels of ethidium uptake in (proteo)liposomes, the ΔpH and $\Delta\psi$ were quantified with the fluorescent probes 2',7'-bis-(2-carboxyethyl)-5-(6)-carboxy-fluorescein (BCECF) and diethyloxycarbocyanine iodide ($\text{DiOC}_2(3)$), respectively. As shown in Fig. 3d, the imposed $-\Delta\text{pH}$ of -55 mV (interior alkaline) dissipated at a rate of $3.2 \pm 0.1 \text{ mV min}^{-1}$ ($n = 3$), whereas the imposed $\Delta\psi$ of -120 mV (interior negative) dissipated at a rate of less than

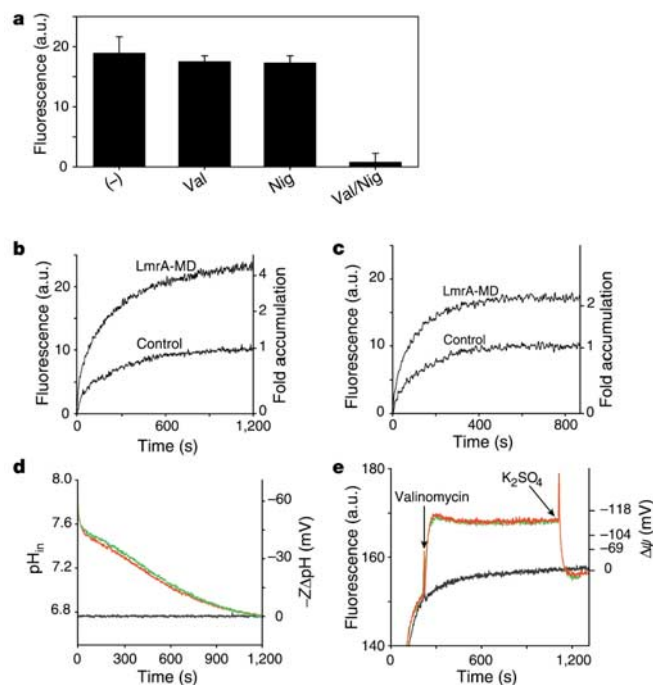


Figure 3 Ethidium transport by LmrA-MD is coupled to Δp . **a**, Effect of ionophore nigericin (Nig) ($\Delta\text{pH} = 0$), valinomycin (Val) ($\Delta\psi = 0$), both (Nig/Val) ($\Delta p = 0$) or neither (—) on LmrA-MD-mediated ethidium uptake in cells ($P < 0.05$, $n = 3$). **b**, **c**, Ethidium uptake in LmrA-MD containing proteoliposomes with imposed ΔpH (interior alkaline) (**b**) or $\Delta\psi$ (interior negative) (**c**). Control, empty liposomes. **d**, **e**, Stability of ΔpH and $\Delta\psi$. Internal pH (pH_{in}) (**d**) and $\text{DiOC}_2(3)$ fluorescence (**e**) in liposomes (red, black) and proteoliposomes (green) with (colour) or without (black) imposed gradients were converted into $-\Delta\text{pH}$ and $\Delta\psi$ values (in mV), respectively.

0.5 mV min^{-1} ($n = 3$) (Fig. 3e). The magnitude and stability of the imposed ΔpH and $\Delta\psi$ were unaffected by the presence of purified LmrA-MD in the liposomal membrane in the absence of ethidium (Fig. 3d, e). These findings indicate that the differences in ethidium uptake in (proteo)liposomes were not a result of the altered partitioning of ethidium in response to protein-induced changes in the imposed electrochemical proton gradient. Instead, the differences in ethidium accumulation are more probably due to drug transport by LmrA-MD directly. As ethidium is a permanently cationic molecule that lacks substituents that can be (de)protonated, LmrA-MD seems to transport ethidium in symport with H^+ .

Evidence for the coupled transport of ethidium and protons was also obtained for ethidium efflux mediated by full-length LmrA in intact *L. lactis* cells. Measurement of the intracellular pH (pH_{in}) in LmrA-expressing cells by using the fluorescent pH probe carboxy-fluorescein succinimidyl ester (CFSE) revealed a primary increase in pH_{in} of $0.17 \pm 0.01 \text{ min}^{-1}$ ($n = 3$) after the addition of L-arginine as a source of metabolic energy, which was comparable to the primary increase of $0.17 \pm 0.02 \text{ min}^{-1}$ ($n = 3$) in non-expressing control cells, and which reflected the generation of a ΔpH (interior alkaline) through proton pumping by the $\text{F}_0\text{F}_1 \text{ H}^+$ -ATPase. The primary increase in pH_{in} was followed by a secondary increase of $(4.95 \pm 0.75) \times 10^{-3} \text{ min}^{-1}$ ($n = 3$) in LmrA-expressing cells in the presence of $5 \mu\text{M}$ ethidium (Fig. 4a, b). In contrast, pH_{in} decreased by $(0.13 \pm 0.10) \times 10^{-3} \text{ min}^{-1}$ ($n = 3$) in these cells in the absence of ethidium (Fig. 4b). The secondary increase in pH_{in} was also absent in control cells or cells expressing a mutant LmrA containing a glutamate $\rightarrow$ alanine substitution at position 314 (E314A) (Fig. 4b). This mutant extruded ethidium at a reduced rate compared with LmrA (Fig. 4c). In addition, E314A mutant LmrA-MD was unable to mediate ΔpH -driven, proton-coupled ethidium transport in proteoliposomes (data not shown). Consistent with these data were measurements of $\text{DiOC}_2(3)$ fluorescence

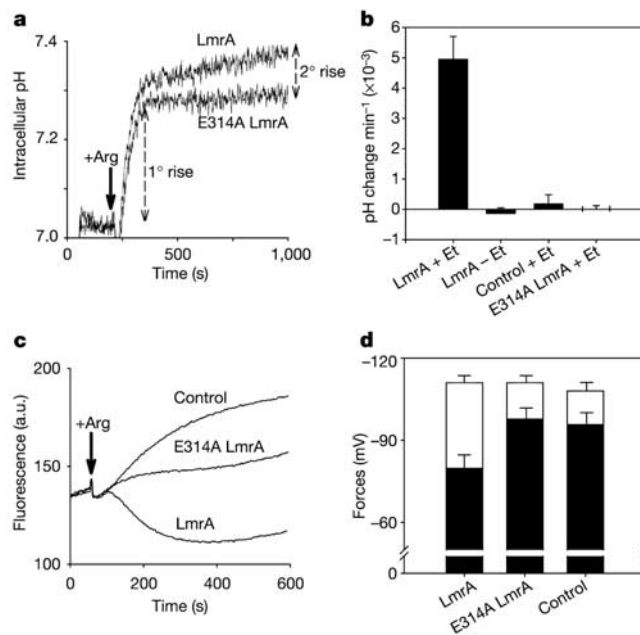


Figure 4 Coupled efflux of ethidium and protons by LmrA in *L. lactis* cells.

a, Measurement of pH_{in} in cells in the presence of $5 \mu\text{M}$ ethidium. Cells expressing E314A mutant LmrA are identical with non-expressing control (not shown). Arg, 20 mM L-arginine. 1°, primary; 2°, secondary. **b**, Slope of the secondary increase in pH_{in} (see **a**) ($P < 0.05$, $n = 3$). **c**, Ethidium efflux. **d**, Measurement of the Δp (interior negative and alkaline) in cells in the presence of ethidium ($P < 0.05$, $n = 3$). Black and white bars show $\Delta\psi$ and $-\Delta\text{pH}$, respectively, the sum of which equals the Δp .

showing that LmrA-expressing cells had a significantly higher $-\Delta p\text{H}$ and lower $\Delta\psi$ than control cells or cells expressing the E314A mutant LmrA in the presence of ethidium (Fig. 4d). The composition of the Δp was equal between the cells without ethidium (data not shown). Thus, the ethidium-dependent increase in $\Delta p\text{H}$ was observed in LmrA-expressing cells by two independent methods.

The activity of primary proton pumps in bacteria is known to be constrained by the Δp , especially by its $\Delta\psi$ component which is generated more readily than the $\Delta p\text{H}$ owing to the buffer capacity of the cytoplasm¹³. The controlled dissipation of excessive $\Delta\psi$ through the activity of electrogenic K^+ uptake systems under these conditions is well established in bacteria, and allows the formation of a $\Delta p\text{H}$ during proton efflux^{14,15}. The higher $\Delta p\text{H}$ and lower $\Delta\psi$ in LmrA-expressing cells in the presence of ethidium therefore indicate enhanced proton extrusion, which is consistent with the notion that LmrA mediates the efflux of ethidium together with protons.

Our results show that LmrA-MD mediates the Δp (interior negative and alkaline)-driven uptake of ethidium in cells, which contain the protein in a right-side-out orientation, and in proteoliposomes, which contain the protein in an inside-out orientation. The membrane domain of primary-active LmrA is therefore functional as a reversible secondary-active drug transporter, acting independently of ATP hydrolysis. In full-length LmrA, ATP hydrolysis by the NBD imposes directionality on transport by the MD and drives the efflux of ethidium and protons from cells against the inwardly directed Δp (interior negative and alkaline). Our observations are consistent with previous, unexplained results indicating that P-glycoprotein might mediate the outward movement of protons in mammalian cells^{16–19}. They also show analogy to observations reported for the *Escherichia coli* ArsAB ATPase, which exhibits a dual mode of energy coupling to arsenite/antimonite efflux depending on the subunit composition²⁰. The remarkable presence of a secondary-active multidrug translocator domain in primary-active LmrA indicates that this multidrug pump might have evolved through the association of a catalytic module and a translocator module, to produce a transporter with altered functionality. □

Methods

Genetic manipulations

To express the N-terminally His₆-tagged LmrA-MD, the corresponding region of the *lmrA* gene was PCR amplified from pNHLmrA²¹, using primers 5'-GGAGGCACTCACCATGG GGC-3' and 5'-CTAGCTAGTCTAGATTATTAATGGTGAGCAGAGAGGCTCTTCC TTCCAAATCAAGACTATAC-3' to introduce a TAA stop codon after His 353 and an *Xba*I site at the 3' end, and cloned as an *Nco*I–*Xba*I fragment in pNZ8048 (ref. 21), yielding pNHLmrA-MD. The E314A mutants LmrA and LmrA-MD were generated with the QuikChange kit (Stratagene) with primers 5'-GGCAACTTCTTTACAGCGCTTG CCAAGCG-3' and 5'-CGCTTTGGCAAGCGCTGTAAAGAAAGTTGCC-3'.

Cell culture and membrane preparations

Growth of *L. lactis* NZ9000 in M17 medium (Difco), the induction of LmrA or LmrA-MD expression using nisin A, and the preparation of inside-out and right-side-out membrane vesicles were performed as described^{6,22}.

Cytotoxicity assays

Drug toxicity was measured as described²³, in the presence of 10 μM vinblastine or 1 μM verapamil when appropriate. The relative drug sensitivity (expressed as a percentage) of LmrA-MD-expressing cells in Fig. 2c was calculated from the drug concentration required for half-maximum inhibition of the growth rate of LmrA-MD-expressing cells (IC_{50}) divided by the IC_{50} for control cells.

Drug transport in cells

Ethidium transport in intact cells was measured by fluorimetry at a concentration of 2 μM ethidium bromide^{22,23}. Initial ethidium transport rates were determined over the first 120 s, during which uptake was linear.

$\Delta p\text{H}$ and $\Delta\psi$ measurements in cells

The pH_{in} in cells was measured by fluorimetry with 5 (and 6-) carboxyfluorescein diacetate succinimidyl ester (CFDASE) (Molecular Probes)²⁴. Metabolic energy in *L. lactis* was generated with L-arginine to avoid interference in the pH measurements by lactic acid formed during glucose fermentation. At the end of each measurement, the Δp was

dissipated in the presence of 1 μM valinomycin and nigericin ($\text{pH}_{\text{in}} = \text{pH}_{\text{out}}$). CFSE fluorescence was then calibrated in a pH range of 6–8 with a pH microelectrode. The $\Delta p\text{H}$ was calculated by subtracting pH_{out} from pH_{in} , and multiplied by the conversion factor Z of 59 to express $-\Delta p\text{H}$ in mV. For measurements of the Δp in cells, the $\Delta\psi$ was measured with 10 μM DiOC₂(3) (Molecular Probes)²⁵. DiOC₂(3) fluorescence was calibrated at imposed $\Delta\psi$ values between -59 and -120 mV through 100-fold dilution of cells in buffer containing 2 μM valinomycin and increasing concentrations of K_2SO_4 . The $-\Delta p\text{H}$ was calculated from the increase in $\Delta\psi$ after the addition of 2 μM nigericin, assuming complete interconversion of the $\Delta p\text{H}$ into the $\Delta\psi$ (ref. 10).

Protein crosslinking

Crosslinking of LmrA-MD was performed in the presence of 0.1 mM disuccinimidyl glutarate with inside-out membrane vesicles (100 μg protein) in 100 mM potassium phosphate buffer pH 7.4 in a total reaction volume of 40 μl .

Accessibility of hexahistidine tag

Membrane vesicles (40 μg protein) in 50 mM K-HEPES buffer (pH 7.4) were incubated with Protease K. The uncleaved His₆ tag was detected on an immunoblot by using the monoclonal anti-pentahistidine antibody (Qiagen).

Photoaffinity labelling

LmrA or LmrA-MD in inside-out membrane vesicles (90 μg protein) was crosslinked with 0.25 μM [3H]azidopine (51 Ci mmol⁻¹; Amersham Biosciences) for 2.5 min by irradiation with ultraviolet light (312 nm). Samples were analysed by SDS-PAGE and autoradiography.

DNA-loaded proteoliposomes

LmrA-MD was purified and functionally reconstituted in proteoliposomes, as described in ref. 22, in 20 mM potassium phosphate buffer pH 6.8 containing 100 mM potassium acetate and 1 mg ml⁻¹ sonicated calf thymus DNA. Liposomes were destabilized by the addition of 3 mM Triton X-100 to allow the unidirectional reconstitution of purified LmrA-MD.

Ethidium transport driven by $\Delta p\text{H}$ and $\Delta\psi$ in proteoliposomes

To impose diffusion gradients, proteoliposomes in 20 mM potassium phosphate buffer pH 6.8 containing 2 mM MgSO_4 , 100 mM potassium acetate and 10 nmol valinomycin per mg protein were diluted 1:100 (to 10 μg protein per ml) into the same buffer (no gradient), into 20 mM potassium phosphate buffer pH 6.8 containing 2 mM MgSO_4 and 50 mM K_2SO_4 to generate a $\Delta p\text{H}$ (interior alkaline), and into 20 mM *N*-methyl-*D*-glucosamine (NMG) phosphate buffer pH 6.8 containing 2 mM MgSO_4 and 100 mM NMG acetate to generate a $\Delta\psi$ (interior negative)²⁶. After 20 s, 0.5 μM ethidium bromide was added and the fluorescence of the ethidium–DNA complex formed inside the proteoliposomes was followed over time, and calibrated with standard ethidium solutions. Initial ethidium transport rates were determined over the first 10 s, during which uptake was linear. For measurements of the imposed $\Delta p\text{H}$, (proteo)liposomes were prepared in buffers containing 100 μM BCECF (Molecular Probes). To measure the $\Delta\psi$, 10 μM DiOC₂(3) was added to diluted proteoliposomes before the addition of valinomycin. BCECF and DiOC₂(3) fluorescence were calibrated as described for $\Delta p\text{H}$ and $\Delta\psi$ measurements.

Data analysis

All statistical analyses were performed with the paired Student's *t*-test with a 95% confidence interval for the sample mean. Where indicated, *n* refers to the number of independent observations with different cell batches and (proteo)liposome preparations.

Received 30 July; accepted 21 October 2003; doi:10.1038/nature02173.

- Ambudkar, S. V. et al. Biochemical, cellular and pharmacological aspects of the multidrug transporter. *Annu. Rev. Pharmacol. Toxicol.* **39**, 361–398 (1999).
- Higgins, C. F. ABC transporters: from microorganisms to man. *Annu. Rev. Cell Biol.* **8**, 67–113 (1992).
- Paulsen, I. T., Brown, M. H. & Skurray, R. A. Proton-dependent multidrug efflux systems. *Microbiol. Rev.* **60**, 575–608 (1996).
- Van Veen, H. W. et al. Multidrug resistance mediated by a bacterial homolog of the human drug transporter MDR1. *Proc. Natl Acad. Sci. USA* **93**, 10668–10672 (1996).
- Van Veen, H. W. et al. A bacterial antibiotic-resistance gene that complements the human multidrug-resistance P-glycoprotein gene. *Nature* **391**, 291–295 (1998).
- Van Veen, H. W. et al. The homodimeric ATP-binding cassette transporter LmrA mediates multidrug transport by an alternating two-site (two-cylinder engine) mechanism. *EMBO J.* **19**, 2503–2514 (2000).
- Chang, G. & Roth, C. B. The structure of MsbA from *E. coli*: a homolog of the multidrug resistance ATP binding cassette (ABC) transporters. *Science* **293**, 1793–1800 (2001).
- Chang, G. Structure of MsbA from *Vibrio cholera*: a multidrug resistance ABC transporter homolog in a closed conformation. *J. Mol. Biol.* **330**, 419–430 (2003).
- Putman, M., Van Veen, H. W., Degener, J. E. & Konings, W. N. Antibiotic resistance: era of the multidrug pump. *Mol. Microbiol.* **36**, 772–774 (2000).
- Kaback, H. R. In *The Bacteria* Vol. 12 (*Bacterial Energetics*) (ed. Krulwich, T. A.) 151–193 (Academic, London, 1990).
- Sharom, F. J., Yu, X. & Doige, C. A. Functional reconstitution of drug transport and ATPase activity in proteoliposomes containing partially purified P-glycoprotein. *J. Biol. Chem.* **268**, 24197–24202 (1993).
- Eytan, G. D., Regev, R. & Assaraf, Y. G. Functional reconstitution of P-glycoprotein reveals an apparent near stoichiometric drug transport to ATP hydrolysis. *J. Biol. Chem.* **271**, 3172–3178 (1996).
- Booth, I. R. Regulation of cytoplasmic pH in bacteria. *Microbiol. Rev.* **49**, 359–378 (1985).
- Kroll, R. G. & Booth, I. R. The relationship between intracellular pH, the pH gradient and potassium transport in *Escherichia coli*. *Biochem. J.* **216**, 709–716 (1983).

15. Bakker, E. P. *Alkali Cation Transport Systems in Prokaryotes* (CRC Press, Florida, 1993).
16. Thiebaut, F. *et al.* Activity of the multidrug transporter results in alkalization of the cytosol: measurement of cytosolic pH by microinjection of a pH-sensitive dye. *J. Histochem. Cytochem.* **38**, 685–690 (1990).
17. Landwojtowicz, E., Nervi, P. & Seelig, A. Real-time monitoring of P-glycoprotein activation in living cells. *Biochemistry* **41**, 8050–8057 (2002).
18. Santa, C. T., Fritz, F. & Roepe, P. D. Effects of ion gradients on H⁺ transport mediated by human MDR1 protein. *Biochemistry* **38**, 4227–4234 (1999).
19. Howard, E. M. & Roepe, P. D. Purified human MDR1 modulates membrane potential in reconstituted proteoliposomes. *Biochemistry* **42**, 3544–3555 (2003).
20. Kuroda, M., Dey, S., Sanders, O. I. & Rosen, B. P. Alternate energy coupling of ArsB, the membrane subunit of the Ars anion-translocating ATPase. *J. Biol. Chem.* **272**, 326–331 (1997).
21. Margolles, A., Putman, M., Van Veen, H. W. & Konings, W. N. The purified and functionally reconstituted multidrug transporter LmrA of *Lactococcus lactis* mediates the transbilayer movement of specific fluorescent phospholipids. *Biochemistry* **38**, 16298–16306 (1999).
22. Reuter, G. *et al.* The ATP-binding cassette multidrug transporter LmrA and lipid transporter MsaA have overlapping substrate specificities. *J. Biol. Chem.* **278**, 35193–35198 (2003).
23. Janvilisri, T. *et al.* Sterol transport by the human breast cancer resistance protein (ABCG2) expressed in *Lactococcus lactis*. *J. Biol. Chem.* **278**, 20645–20651 (2003).
24. Breeuwer, P., Drocourt, J., Rombouts, F. M. & Abee, T. A novel method for continuous determination of the intracellular pH in bacteria with the internally conjugated fluorescent probe 5 (and 6-) carboxyfluorescein succinimidyl ester. *Appl. Environ. Microbiol.* **62**, 178–183 (1996).
25. Novo, D., Perlmutter, N. G., Hunt, R. H. & Shapiro, H. M. Accurate flow cytometric membrane potential measurement in bacteria using diethyloxycarbocyanine and a ratiometric technique. *Cytometry* **35**, 55–63 (1999).
26. Van Veen, H. W., Abee, T., Kortstee, G. J. J., Konings, W. N. & Zehnder, A. J. B. Mechanism and energetics of the secondary phosphate transport system of *Acinetobacter johnsonii* 210A. *J. Biol. Chem.* **268**, 19377–19383 (1993).

Acknowledgements We thank Chris Higgins, Peter McNaughton, Ben Luisi, and Ian Booth for stimulating discussions. This research was funded by Cancer Research UK, the Association of International Cancer Research (AICR), the Biotechnology and Biological Sciences Research Council (BBSRC), the Medical Research Council (MRC), the Royal Society, and Molecular Devices Ltd. S.V. was the recipient of a Cambridge Nehru Scholarship.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to H.W.v.V. (hvw20@cam.ac.uk).

The Bloom's syndrome helicase suppresses crossing over during homologous recombination

Leonard Wu & Ian D. Hickson

Cancer Research UK, Weatherall Institute of Molecular Medicine, University of Oxford, John Radcliffe Hospital, Oxford, OX3 9DS, UK

Mutations in *BLM*, which encodes a RecQ helicase, give rise to Bloom's syndrome, a disorder associated with cancer predisposition and genomic instability¹. A defining feature of Bloom's syndrome is an elevated frequency of sister chromatid exchanges². These arise from crossing over of chromatid arms during homologous recombination, a ubiquitous process that exists to repair DNA double-stranded breaks and damaged replication forks. Whereas crossing over is required in meiosis, in mitotic cells it can be associated with detrimental loss of heterozygosity. *BLM* forms an evolutionarily conserved complex with human topoisomerase III α (hTOPO III α)^{3,4}, which can break and rejoin DNA to alter its topology. Inactivation of homologues of either protein leads to hyper-recombination in unicellular organisms⁵. Here, we show that *BLM* and hTOPO III α together effect the resolution of a recombination intermediate containing a double Holliday junction. The mechanism, which we term double-junction dissolution, is distinct from classical Holliday junction resolution and prevents exchange of flanking sequences. Loss of such an activity explains many of the cellular

phenotypes of Bloom's syndrome. These results have wider implications for our understanding of the process of homologous recombination and the mechanisms that exist to prevent tumorigenesis.

Previous analyses have revealed that RecQ helicases can stimulate the DNA strand passage activity of type IA topoisomerases. Specifically, the *Escherichia coli* RecQ protein has been shown to stimulate *E. coli* Topo III to catenate supercoiled plasmids⁶. We have shown previously that *BLM* is able to stimulate the ability of hTOPO III α to relax supercoiled DNA, but in those studies we found no evidence of DNA catenation⁷. We surmised that the catenation of supercoiled plasmids is a specific reaction catalysed by the bacterial proteins, and that bacterial and human type IA topoisomerases may have evolved functions that are differentially influenced by the action of RecQ helicases. Given the elevated frequency of recombination in RecQ helicase and type IA topoisomerase mutants, together with the findings that *BLM* can promote branch migration of Holliday junctions (HJs)⁸, and that the sole yeast type IA topoisomerase, Top3, is implicated in the resolution of recombination intermediates⁹, we addressed whether *BLM* and hTOPO III α might act together to process a HJ. In initial studies, we used a molecule that contains a single HJ (X12), which has been characterized previously with respect to both branch migration and resolution by *E. coli* RuvABC^{10,11}. However, we obtained no evidence that hTOPO III α alone (or in combination with *BLM*) could cleave this substrate (Supplementary Fig. 1). Although X12 is a useful mimic of HJs, the torsional constraints on the four arms of a HJ that would be evident *in vivo* are not present in X12, where the arms are short and free to rotate in solution. We postulated that such constraints are required for *BLM* to generate a structure upon which the DNA cleavage activity of hTOPO III α will act. Therefore, we analysed whether *BLM* and hTOPO III α could act upon HJ-containing substrates in which torsional constraints are present. For these studies, we used a previously described oligonucleotide-based structure, here termed DHJ, which contains a double HJ¹². Such structures exist *in vivo* and are thought to arise when both ends of a DNA double-stranded break (DSB) invade an homologous sequence, as proposed in the canonical model of homologous recombination^{13,14}. Whereas RuvC-like resolution activities on single HJs have been identified recently in human cells^{15,16}, the mechanism for resolution of a double HJ is potentially very different: because of their close proximity, neither junction would be expected to be able to adopt the square planar configuration required for branch migration and resolution of the crossover structures¹⁷.

The DHJ substrate comprises two oligonucleotides, B1 and R1, which in all subsequent figures are coloured black and red, respectively (Fig. 1a). Annealing of B1 and R1 results in the formation of two HJs that flank two intervening regions of heteroduplex DNA, each 14 base pairs in length. The ends of each oligonucleotide, one of which is labelled at the 5' end, are juxtaposed in the heteroduplex regions, allowing them to be ligated and thereby allowing circularization of the oligonucleotides (Fig. 1a). This circularization effectively introduces two links between the two circular oligonucleotides, one in each region of heteroduplex DNA (Fig. 1b). Each crossover in DHJ, therefore, has its two heteroduplex arms constrained from free rotation by the other crossover. To confirm the nature of the substrate, unlinking of the circular oligonucleotides was achieved by digestion with the restriction enzymes *HhaI* or *RsaI*, which cut respectively in one of the arms of oligonucleotide B1 and R1 (Fig. 1c, d). This analysis confirmed the duplex nature of the outer arms of DHJ. The products from such a digestion were an intact circle and a linearized oligonucleotide reduced in length respectively by 20 and 16 nucleotides for *RsaI* and *HhaI* digests. These products, when separated on a denaturing gel, had the expected lower and higher mobilities, respectively, compared with the original linear oligonucleotide (Fig. 1c, d). Further

confirmation of the nature of the products of restriction digestion was obtained by determining the sensitivity of the products to degradation by exonuclease I (ExoI); linear products are sensitive to ExoI treatment whereas circular products are resistant (Fig. 1c, d).

We examined whether BLM and/or hTOPO III α could act upon DHJ. To aid interpretation of these analyses, all reaction products were run on denaturing gels to remove any potentially confounding effects of DNA secondary structure on electrophoretic mobility. With this system, we could assess confidently any changes in topology or in molecular mass of the labelled oligonucleotide. BLM or hTOPO III α alone had no effect on the size or topological state of DHJ (Fig. 2a). However, incubation of both proteins together with DHJ resulted in a single product (designated P) that co-migrated with the labelled, circular oligonucleotide generated in control reactions by restriction digestion (Fig. 2a). Regardless of whether the label was incorporated into oligonucleotide R1 or B1, the combined action of BLM and hTOPO III α released a labelled product that co-migrated with circular R1 or B1 oligonucleotides (data not shown). This indicates that the reaction is symmetrical and results in the conversion of DHJ into two equally-sized products. The precise structure of these products was ascertained in experiments described below.

We next determined whether the reaction catalysed by BLM and hTOPO III α on DHJ required the known catalytic activities of both proteins. For this, we used a mutant form of hTOPO III α , hTOPO III α (Y337F), in which the invariant tyrosine residue in the active site had been mutated to phenylalanine¹⁸. Consistent with a require-

ment for the strand passage activity of hTOPO III α , the hTOPO III α (Y337F) mutant was not able to support the generation of any products (Fig. 2b). The helicase activity of BLM is strictly dependent on ATP hydrolysis¹⁹. No products were observed when the non-hydrolysable ATP analogue AMP-PNP was substituted for ATP in the reaction, indicating that the ATPase activity, and hence the helicase activity, of BLM is required for the generation of circular products (Fig. 2c). Owing to the covalently closed nature of DHJ, which is resistant to boiling, we were not able to address directly whether BLM could unwind DHJ (Fig. 2d). However, analysis under native gel conditions revealed that BLM was able to unwind DHJ in which oligonucleotide R1 had been linearized by *RsaI* digestion, indicating that BLM can both bind to and catalyse the unwinding of DHJ (Fig. 2d).

We then investigated whether other proteins with similar catalytic activities could be substituted for BLM or hTOPO III α . Eukaryotic topoisomerase I was not able to support the generation of any products, even when a vast excess of active topoisomerase I units were added to reactions containing BLM (Fig. 3a). Similarly, when BLM was substituted with another helicase, *E. coli* UvrD, no products were observed, even though UvrD was able to efficiently disrupt the *RsaI*-digested DHJ molecule (Fig. 3b). We conclude that the products generated from DHJ are specific for the combination of BLM and hTOPO III α .

There are two straightforward models for how BLM and hTOPO III α might act together to convert DHJ into two equally sized products. First, BLM and hTOPO III α might be promoting classical

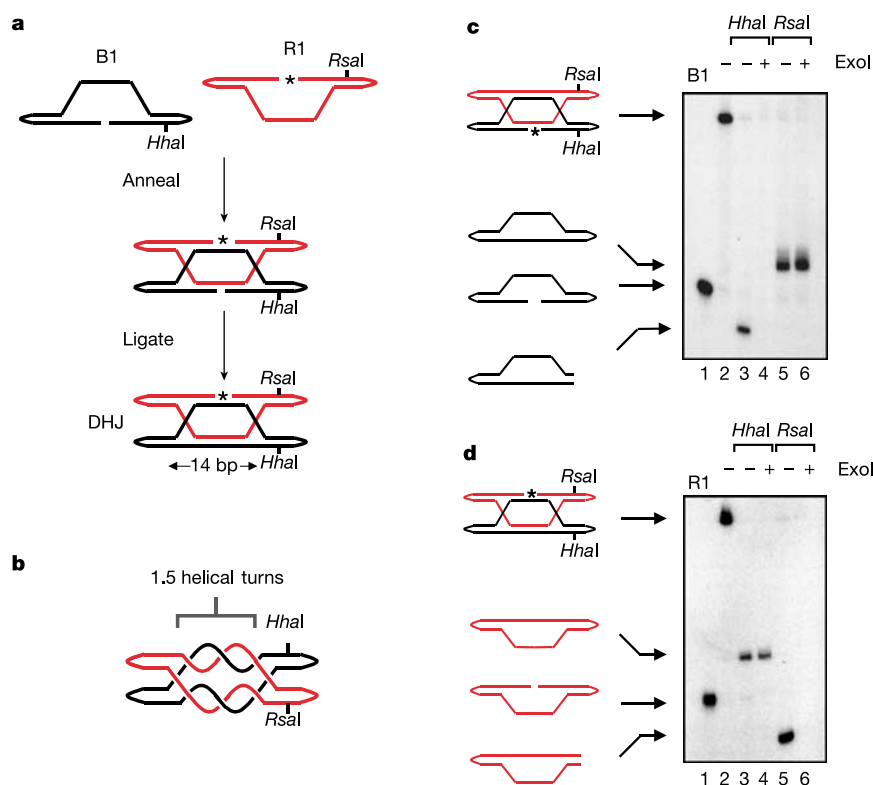


Figure 1 Generation of a substrate containing a double HJ. **a**, Schematic representation of the oligonucleotides and steps used in the construction of DHJ. **b**, Schematic representation of strand intertwinings present in DHJ. Only those intertwinings that result in the formation of a topological link between oligonucleotides B1 and R1 are shown. For clarity, DHJ is depicted in subsequent figures as it appears in **a**. **c**, 8% denaturing polyacrylamide gel electrophoresis (PAGE) confirming the structure of DHJ in which oligonucleotide B1 is labelled (indicated by the asterisk). Representations of the label in

different molecules and their relative electrophoretic mobilities are shown. For clarity all potential DNA secondary structure is shown in this and subsequent figures. Lane 1, labelled B1 oligonucleotide; lane 2, DHJ; lanes 3, 4, DHJ digested with *HhaI*; lanes 5, 6, DHJ digested with *RsaI*. The products in lanes 4 and 6 were treated with ExoI as indicated above. **d**, 8% denaturing PAGE confirming structure of DHJ in which oligonucleotide R1 was labelled. Lane designations are as in **c**.

HJ resolution, analogous to the reaction catalysed by *E. coli* RuvC. This would entail the symmetrical nicking of two strands at the crossover points of each of the junctions, thereby releasing two identically sized duplexes¹⁷. As HJs can be resolved in one of two orientations, released duplexes would be identical in size but their sequence would depend on the orientation of resolution of each of the junctions (Fig. 4a). In the second model, BLM and hTOPO III α might generate products by catalysing two specific strand passage events, one in each of the heteroduplex regions between the junctions, thereby unlinking circular B1 and R1 oligonucleotides (Fig. 4a). To distinguish between these mechanistically distinct modes of DHJ resolution, we analysed, in more detail, the structure of the products of the coupled reaction. In particular, we followed the fate of the outer arms of DHJ. The right-hand arms were distinguished using the restriction enzymes *HhaI* and *RsaI*, which cut oligonucleotides B1 and R1, respectively. The left-hand arms were differentiated by the addition of an internal biotin moiety to the hairpin loop of either B1 or R1. If BLM and hTOPO III α were catalysing classical, RuvC-like HJ resolution, then analysis of the products should reveal that the restriction site and biotin markers segregate with the label approximately 50% of the time, indicating an exchange of flanking arms (Fig. 4a). In contrast, the strand passage model predicts that DHJ would be separated into its two constituent oligonucleotides and, therefore, that there would be no exchange of arms that flank the two HJs (Fig. 4a).

Figure 4b shows the result of such an analysis of products generated from DHJ in which B1 was radiolabelled and the biotin moiety was attached to R1. Comparing lanes 2 and 3 confirms that, in this variant of DHJ, the label on B1 was covalently associated with the biotin group, as it could be depleted by streptavidin affinity beads, even after the substrate had been denatured by boiling. In contrast, the labelled product generated by the combined actions of BLM and hTOPO III α was resistant to depletion by streptavidin (compare lanes 4 and 5). This indicates that the left arm of the labelled product was derived exclusively from B1. Digestion of the labelled BLM–hTOPO III α product with *HhaI* or *RsaI* revealed that it was completely sensitive to *HhaI* digestion, but was resistant to *RsaI* digestion (compare lanes 4, 6 and 7), indicating that the right arm of the labelled product was also derived exclusively from B1. These data are consistent with the generation of products with no exchange of the arms that flank the HJs. The same conclusion was arrived at from analysis of the reciprocal DHJ molecule in which R1 was labelled and the biotin moiety was attached to B1 (data not shown). We conclude that BLM and hTOPO III α can convert DHJ into two equally-sized products through a strand passage mechanism that decatenates the intact circular B1 and R1 oligonucleotides, and not through classical HJ resolution. Moreover, although HJ resolution can occur in one of two orientations, the reaction catalysed by BLM and hTOPO III α has a single outcome; that is, the separation of the recombining molecules without exchange of the

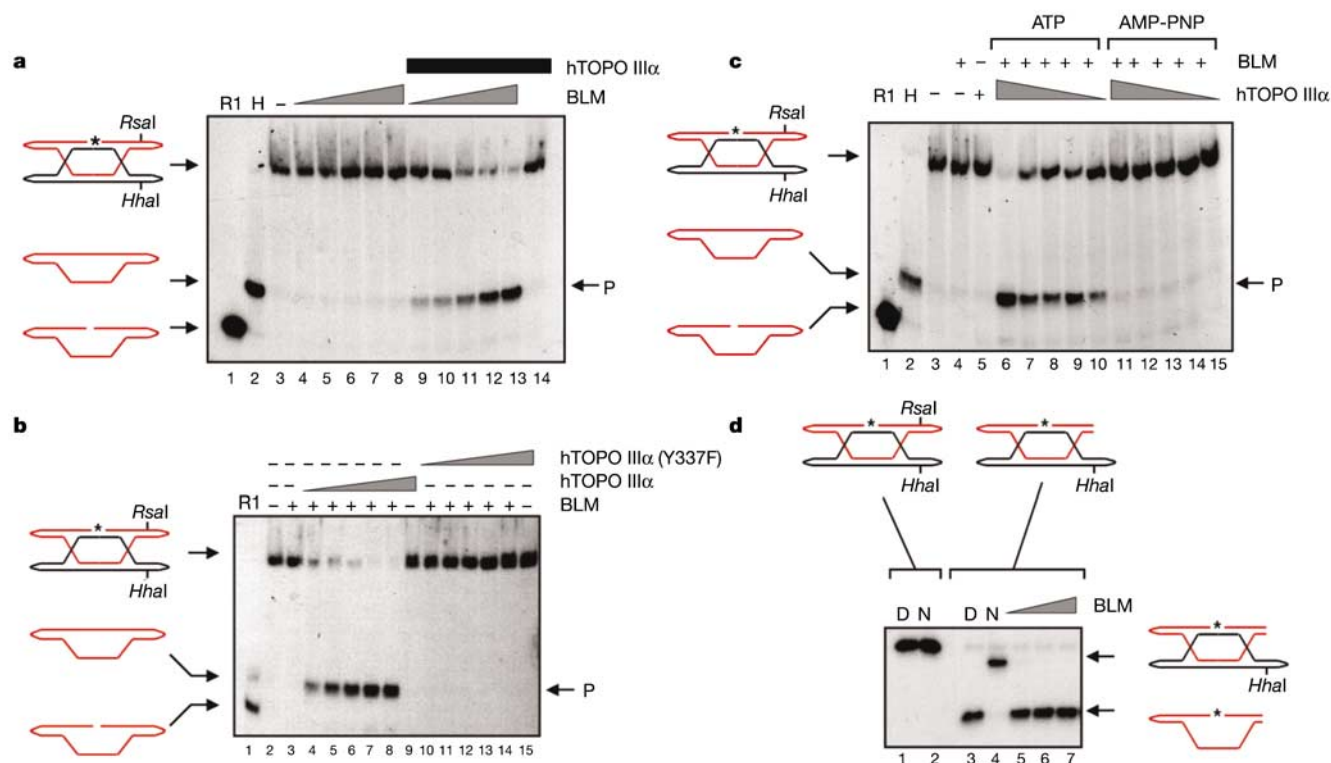


Figure 2 BLM and hTOPO III α can convert DHJ into circular products. **a**, DHJ was incubated alone (lane 3) or with increasing concentrations of BLM (0.75–12 nM; lanes 4–8 and 9–13), as indicated. A total of 250 nM hTOPO III α was included in lanes 9–14. Labelled linear R1 and circular R1, as released by *HhaI* digestion of DHJ, were run as markers in lanes 1 and 2, respectively. The position of product P is indicated. **b**, The reaction requires catalytically active hTOPO III α . DHJ was incubated alone (lane 2) or with increasing concentrations (15–250 nM) of either hTOPO III α (lanes 4–9) or hTOPO III α (Y337F) (lanes 10–15), as indicated. A total of 6 nM BLM was included in lanes 3–8 and 10–14, as indicated. Labelled linear R1 and circular R1 were run together in lane 1 as markers. **c**, Generation of product P requires the ATPase activity of BLM. DHJ was

incubated alone (lane 3) or with 6 nM BLM (lanes 4 and 6–15) or 250 nM hTOPO III α alone (lane 5). In lanes 6–10 and 11–15, hTOPO III α was also included at concentrations ranging from 15 to 250 nM, as indicated. A total of 5 mM ATP or AMP-PNP was included in lanes 6–10 and 11–15, respectively. Lanes 1 and 2 contain markers as above. **d**, BLM can catalyse unwinding of DHJ. DHJ or *RsaI*-digested DHJ was run on a native gel. Material in lanes 1 and 3 was denatured (D). Native (N) DHJ and *RsaI*-digested DHJ were run in lanes 2 and 4, respectively. BLM was included at a final concentration of 1.5 nM (lane 5), 3 nM (lane 6) or 6 nM (lane 7). Representations of the labelled molecules are indicated.

genetic material that flanks the crossover points. We refer to this mechanism as double-junction dissolution in order to distinguish it from classical HJ resolution.

The formation of crossover and non-crossover products during homologous recombination has been proposed to occur through the use of independent pathways, as opposed to two alternative outcomes of a single reaction²⁰. *In vivo*, these pathways are highly regulated, with the formation of crossovers being promoted during meiosis but suppressed during mitosis²¹. Crossovers are thought to arise from a RuvC-like resolution of double HJs but can also be generated by the sequential nicking of two single HJs that arise from the first- and second-end capture steps of the canonical model of DSB repair²². In contrast, non-crossovers are thought to arise through synthesis-dependent strand annealing²³. We propose that double-junction dissolution catalysed by BLM and hTOPO III α

provides a newly discovered mechanism by which homologous recombination repair events can give rise to non-crossover products. Indeed, recent genetic evidence has revealed that the yeast homologues of BLM and hTOPO III α , Sgs1 and Top3 respectively, function in a pathway distinct from synthesis-dependent strand annealing to suppress the formation of crossover products arising from homologous recombination repair²⁴. The model proposed by ref. 24 to explain the role of Sgs1p and Top3p in crossover suppression was that these proteins act on recombination intermediates containing double HJs. Similar models involving topoisomerase-mediated processing of double HJs into non-crossover products have also been suggested²⁵. Our biochemical data provide a direct demonstration of the validity of such models.

Although double HJs have been found *in vivo* during meiotic DSB repair, their existence during mitotic DSB repair remains to be confirmed. It is possible that double HJs form rarely during mitotic DSB repair and that most of the repair events occur via synthesis-dependent strand annealing. Indeed, Bloom's syndrome cells are not

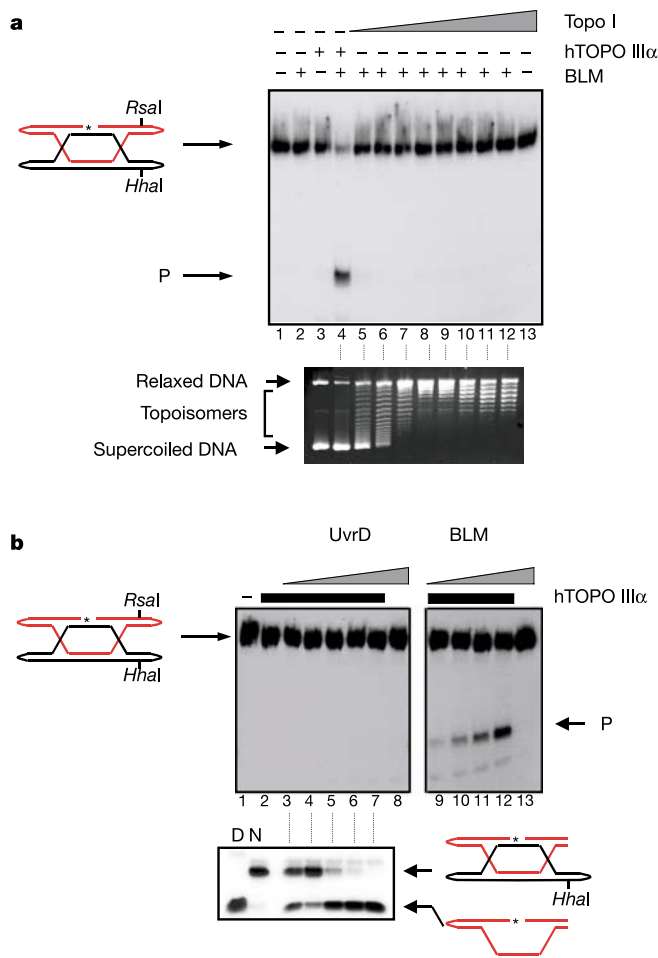


Figure 3 The generation of circular products is specific for the combination of BLM and hTOPO III α . **a**, Top panel: DHJ was incubated alone (lane 1), or with increasing concentrations of wheat germ topoisomerase I (from 0.07 to 10 U in lanes 5–13), as indicated, or with 250 nM hTOPO III α (lanes 3 and 4). A total of 6 nM BLM was included in lanes 2 and lanes 4–12, as indicated. Bottom panel: DNA relaxation assay on supercoiled Φ X174 DNA. The DNA was untreated (lane 3) or incubated with topoisomerase III α or I at equivalent concentrations to those corresponding to the lanes of the top panel (vertical lines link panels). **b**, Top panel: DHJ was incubated alone (lane 1) or with increasing concentrations of UvrD from 1.85 to 150 nM (lanes 3–8) or BLM from 0.75 to 6 nM (lanes 9–13). A total of 250 nM hTOPO III α was included in lanes 2–7 and 9–12, as indicated. Bottom panel: helicase assay on *RsaI*-digested DHJ. Lanes D and N contain the substrate after or before denaturation, respectively. UvrD was included in lanes 3–7 at concentrations equivalent to those in the corresponding lanes in the upper panel (vertical lines link panels). Representations of the labelled DNA molecules are indicated.

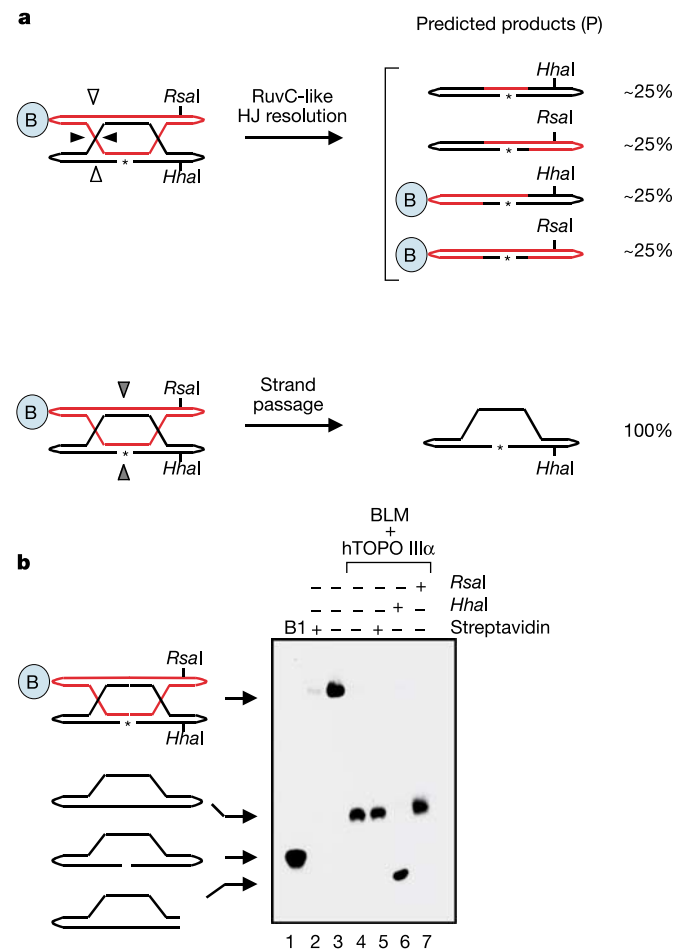


Figure 4 Analysis of the reaction mechanism for the generation of circular products. **a**, Schematic representation of the two possible modes of DHJ processing and the predicted products. An internal biotin moiety is shown as a blue circle containing a B. Black and white arrowheads indicate the two possible orientations of resolution of each HJ within DHJ. For simplicity, arrowheads have only been drawn on one of the HJs. Grey arrowheads indicate proposed positions of strand passage events. For both models, only labelled reaction products are shown, together with their predicted relative abundance (see text for details). **b**, Analysis of the arms of DHJ resolution products. DHJ consisting of labelled B1 and biotinylated R1 was untreated (lane 3) or subjected to depletion by streptavidin affinity beads (lane 2). DHJ reaction products, generated by BLM and hTOPO III α , were analysed after various treatments, as indicated (lanes 4–7). Lane 1 contains labelled linear B1.

hypersensitive to ionizing radiation, indicating that BLM does not have an essential role in general DSB repair²⁶. Rather, we propose that double HJs are formed during the homologous recombination-dependent repair of daughter strand gaps that arise during replication, and that the dissolution of these double HJs by BLM prevents the diagnostically high sister chromatid exchange frequency seen in Bloom's syndrome cells (Supplementary Fig. 2). Furthermore, BLM-catalysed double-junction dissolution may act to suppress tumorigenesis by preventing loss of heterozygosity, a feature associated with BLM deficiency in mice²⁷, through the suppression of ectopic recombination and crossing over between homologous chromosomes²⁸. □

Methods

Protein and DNA reagents

BLM was purified as described previously¹⁹. hTOPO III α and hTOPO III α (Y337F) were gifts from H. Goulaouic and J.-F. Riou. UvrD was provided by S. Matson. Wheat germ topoisomerase I was purchased from Promega. Oligonucleotides and procedures used to generate DHJ were those described previously¹².

DHJ assays

Typically, DHJ (15 fM) was incubated with the indicated proteins in 50 mM Tris-HCl, pH 7.5, 50 mM NaCl, 4 mM MgCl₂, 1 mM dithiothreitol, 0.1 mg ml⁻¹ BSA and, unless otherwise stated, 5 mM ATP, at 37 °C for 60 min. Where indicated, further analysis of reaction products by restriction digestion was performed by the addition of the appropriate enzyme and supplementation of the reaction with 6 mM MgCl₂. Biotin depletion of samples was performed by incubating samples with paramagnetic streptavidin beads (Sigma), and streptavidin-biotin complexes were removed by magnetic retention. Reactions were stopped by the addition 50 mM EDTA and 1% SDS. Samples were de-proteinized by the addition of 100 μ g proteinase K and incubation for a further 15 min at 37 °C. Each sample had an equal volume of 50% urea added to it before boiling and running on an 8% polyacrylamide denaturing gel. Gels were dried onto 3MM paper and exposed to X-ray film.

Topoisomerase and helicase assays

Reaction conditions were those used for DHJ assays, and samples were resolved under native gel conditions.

Received 22 October; accepted 26 November 2003; doi:10.1038/nature02253.

- Ellis, N. A. *et al.* The Bloom's Syndrome gene product is homologous to RecQ helicases. *Cell* **83**, 655–666 (1995).
- Chaganti, R. S., Schonberg, S. & German, J. A. Manyfold increase in sister chromatid exchanges in Bloom's syndrome lymphocytes. *Proc. Natl Acad. Sci. USA* **71**, 4508–4512 (1974).
- Wu, L. *et al.* The Bloom's syndrome gene product interacts with topoisomerase III. *J. Biol. Chem.* **275**, 9636–9644 (2000).
- Johnson, F. B. *et al.* Association of the Bloom syndrome protein with topoisomerase III α in somatic and meiotic cells. *Cancer Res.* **60**, 1162–1167 (2000).
- Wu, L. & Hickson, I. D. RecQ helicases and topoisomerases: components of a conserved complex for the regulation of genetic recombination. *Cell. Mol. Life Sci.* **58**, 894–901 (2001).
- Harmon, F. G., DiGate, R. J. & Kowalczykowski, S. C. RecQ helicase and topoisomerase III comprise a novel DNA strand passage function: a conserved mechanism for control of DNA recombination. *Mol. Cell* **3**, 611–620 (1999).
- Wu, L. & Hickson, I. D. The Bloom's syndrome helicase stimulates the activity of human topoisomerase III α . *Nucleic Acids Res.* **30**, 4823–4829 (2002).
- Karow, J. K., Constantinou, A., Li, J. L., West, S. C. & Hickson, I. D. The Bloom's syndrome gene product promotes branch migration of Holliday junctions. *Proc. Natl Acad. Sci. USA* **97**, 6504–6508 (2000).
- Gangloff, S., de Massy, B., Arthur, L., Rothstein, R. & Fabre, F. The essential role of yeast topoisomerase III in meiosis depends on recombination. *EMBO J.* **18**, 1701–1711 (1999).
- Benson, F. E. & West, S. C. Substrate specificity of the *Escherichia coli* RuvC protein. Resolution of three- and four-stranded recombination intermediates. *J. Biol. Chem.* **269**, 5195–5201 (1994).
- Lloyd, R. G. & Sharples, G. J. Processing of recombination intermediates by the RecG and RuvAB proteins of *Escherichia coli*. *Nucleic Acids Res.* **21**, 1719–1725 (1993).
- Fu, T. T., Tse-Dinh, Y. C. & Seeman, N. C. Holliday junction crossover topology. *J. Mol. Biol.* **236**, 91–105 (1994).
- Schwacha, A. & Kleckner, N. Identification of double Holliday junctions as intermediates in meiotic recombination. *Cell* **83**, 783–791 (1995).
- Szostak, J. W., Orr-Weaver, T. L., Rothstein, R. J. & Stahl, F. W. The double-strand-break repair model for recombination. *Cell* **33**, 25–35 (1983).
- Constantinou, A., Chen, X. B., McGowan, C. H. & West, S. C. Holliday junction resolution in human cells: two junction endonucleases with distinct substrate specificities. *EMBO J.* **21**, 5577–5585 (2002).
- Constantinou, A., Davies, A. A. & West, S. C. Branch migration and Holliday junction resolution catalyzed by activities from mammalian cells. *Cell* **104**, 259–268 (2001).
- West, S. C. Processing of recombination intermediates by the RuvABC proteins. *Annu. Rev. Genet.* **31**, 213–244 (1997).
- Goulaouic, H. *et al.* Purification and characterization of human DNA topoisomerase III α . *Nucleic Acids Res.* **27**, 2443–2450 (1999).
- Karow, J. K., Chakraverty, R. K. & Hickson, I. D. The Bloom's syndrome gene product is a 3'-5' DNA helicase. *J. Biol. Chem.* **272**, 30611–30614 (1997).

- Allers, T. & Lichten, M. Differential timing and control of noncrossover and crossover recombination during meiosis. *Cell* **106**, 47–57 (2001).
- Johnson, R. D. & Jasin, M. Sister chromatid gene conversion is a prominent double-strand break repair pathway in mammalian cells. *EMBO J.* **19**, 3398–3407 (2000).
- Osman, F., Dixon, J., Doe, C. L. & Whitby, M. C. Generating crossovers by resolution of nicked Holliday junctions: a role for Mus81-Eme1 in meiosis. *Mol. Cell* **12**, 761–774 (2003).
- Paques, F. & Haber, J. E. Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.* **63**, 349–404 (1999).
- Ira, G., Malkova, A., Liberi, G., Foiani, M. & Haber, J. E. Srs2 and Sgs1-Top3 suppress crossovers during double-strand break repair in yeast. *Cell* **115**, 401–411 (2003).
- Nasmyth, K. A. Molecular genetics of yeast mating type. *Annu. Rev. Genet.* **16**, 439–500 (1982).
- Bachrati, C. Z. & Hickson, I. D. RecQ helicases: suppressors of tumorigenesis and premature aging. *Biochem. J.* **374**, 577–606 (2003).
- Luo, G. *et al.* Cancer predisposition caused by elevated mitotic recombination in Bloom mice. *Nature Genet.* **26**, 424–429 (2000).
- Hickson, I. D. RecQ helicases: caretakers of the genome. *Nature Rev. Cancer* **3**, 169–178 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank H. Goulaouic, J.-F. Riou and S. Matson for proteins, and members of the Genomic Integrity Group for useful discussions. We also thank J. Haber, S. C. West, E. Louis, P. McHugh and R. Borts for comments on the manuscript, and J. Haber for communicating results before publication. This work is supported by the Cancer Research UK.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to I.D.H. (ian.hickson@cancer.org.uk).

Recognition of small interfering RNA by a viral suppressor of RNA silencing

Keqiong Ye, Lucy Malinina & Dinshaw J. Patel

Structural Biology Program, Memorial Sloan-Kettering Cancer Center, New York, New York 10021, USA

RNA silencing (also known as RNA interference) is a conserved biological response to double-stranded RNA that regulates gene expression, and has evolved in plants as a defence against viruses^{1–3}. The response is mediated by small interfering RNAs (siRNAs), which guide the sequence-specific degradation of cognate messenger RNAs. As a counter-defence, many viruses encode proteins that specifically inhibit the silencing machinery^{3,4}. The p19 protein from the tombusvirus is such a viral suppressor of RNA silencing⁵ and has been shown to bind specifically to siRNA⁶. Here, we report the 1.85-Å crystal structure of p19 bound to a 21-nucleotide siRNA, where the 19-base-pair RNA duplex is cradled within the concave face of a continuous eight-stranded β -sheet, formed across the p19 homodimer interface. Direct and water-mediated intermolecular contacts are restricted to the backbone phosphates and sugar 2'-OH groups, consistent with sequence-independent p19-siRNA recognition. Two α -helical 'reading heads' project from opposite ends of the p19 homodimer and position pairs of tryptophans for stacking over the terminal base pairs, thereby measuring and bracketing both ends of the siRNA duplex. Our structure provides an illustration of siRNA sequestering by a viral protein.

siRNAs, which are the processed products of Dicer⁷ (an RNase III family nuclease), are RNA sequences 21–25 nucleotides (nt) long, composed of duplex (19–23 base pairs, bp) and 3'-overhang (2-nt) segments, with 5'-phosphate and 3'-hydroxyl termini. We first sought to define the siRNA structural elements recognized by the tomato bushy stunt virus (TBSV) p19 protein (Fig. 1a). In agreement with previous research⁶, p19 binds efficiently to a 19-bp siRNA with 3'-dinucleotide overhangs at both ends, but does not recognize

the component single-stranded RNAs (ssRNAs) (Fig. 1c, lanes 1–6). p19 also recognizes duplex RNAs with blunt ends, though less efficiently for 21-bp relative to 19-bp duplexes (Fig. 1c, lanes 7–10). The molecular recognition process distinguishes RNA duplexes from their DNA counterparts, because the corresponding 19-bp DNA duplex with the same sequence does not form the p19 complex (Fig. 1c, lanes 11, 12). We next screened for p19 binding to siRNAs ranging in length from 16–25 bp, and observed optimal binding at the 19–21-bp duplex level (Fig. 1d, lanes 7–12), with gradual attenuation for either shorter (16–18 bp) (Fig. 1d, lanes 1–6) or longer (22–25 bp) (Fig. 1d, lanes 13–20) duplexes. This establishes that p19 recognizes RNA duplexes of a defined length, without the requirement of 3'-dinucleotide overhangs.

We have determined the structure of homodimeric p19 (residues 27–158) in complex with a 21-nt (19-bp) siRNA (Fig. 1b) by X-ray crystallography at 1.85-Å resolution (stereo view, Fig. 2a). The concave saddle-like β -sheet surface of the p19 dimer spans all 19 bp along one face of the siRNA duplex (Fig. 2b, c), thereby burying 1,600 Å² (22%) of solvent-accessible surface area of the bound RNA. Both ends of the 19-bp RNA duplex are bracketed by the projecting helices H1 and H1' of the bound p19 homodimer (Fig. 2a). Intermolecular protein–RNA contacts span the entire length of the p19 homodimer surface (green regions in Fig. 2d), with the dimer surface electrostatically positively charged (blue regions, Fig. 2e) for those patches that contact the RNA major groove.

Each monomer within the p19 homodimer contains four α -helices (H1–H4) and four β -strands (S1–S4) (Fig. 3a). The protein dimerizes through pairing of antiparallel β -strands (S4–S4') and antiparallel α -helices (H4–H4'), thereby burying 1,300 Å² of total solvent-accessible area. The homodimer adopts basically a rectangular two-layered structure from which project two symmetry-related H1 and H1' (reading-head) helices. One layer (concave side) is comprised of a strongly curved eight-stranded antiparallel β -sheet (Fig. 3a), while the other layer consists of six helices packed on the convex side. The symmetry-related 'reading-head' helices H1 and H1' are anchored through conserved side-chain interactions to the core β -sheet scaffold of the p19 homodimer (Fig. 3b).

The non-symmetric siRNA duplex is crossed by a crystallographic dyad in the crystal (see Methods), resulting in two opposite orientations of the siRNA duplex in the complex. These two orientations are shown in Fig. 3c, together with colour-coding for temperature factors: blue (low) graded to red (high). The extent of mobility increases on proceeding from the protein-contacting to the solvent-exposed segments of the bound siRNA. The RNA helix basically adopts an A-form conformation, with the helix axis bending by $\sim 40^\circ$ towards the protein (Fig. 3c). The 3' dinucleotide overhangs of the bound siRNA are not visible in the electron density map, indicating that they are disordered and not recognized by the p19 homodimer in the complex. This is consistent with our mobility shift experiments (Fig. 2c) where blunt-end duplexes bind p19 as well as their 3' overhang-containing counterparts. Nevertheless, the 3' overhangs may be important for other aspects of siRNA recognition and function⁸.

The p19 homodimer monitors the length of the siRNA duplex in the structure of the complex. It achieves this unique caliper-like capability by bracketing both ends of the duplex with a pair of tryptophans, which project off the amino-terminal ends of symmetry-related 'reading head' H1 and H1' helices (Fig. 2a). Stacking between W39 with the 3'-terminal base and W42 with the 5'-terminal base essentially extends the terminal base pairs at either end by an additional step (Fig. 4a). Such bracketing by pairs of stacked tryptophans at either end is further anchored by the stacking of the long side chain of R43 over W39 (Fig. 4a). Importantly, W42 is universally conserved among all available p19 homologues, while W39 can be replaced by leucine, arginine or serine residues (Fig. 1a), suggestive of variable recognition of the

3'-end of the RNA duplex. There are also direct and water-mediated intermolecular contacts between backbone and side chains of amino acids that project from the β -strands (S2) and the ordered N-terminal segments, with the phosphate backbone and sugar 2'-hydroxyls positioned towards the RNA 5'-end (Fig. 4a). The ability of p19 to accommodate duplexes ranging from 19 to 21 bp could originate in structural plasticity of the distance separating 'reading head' helices, each of which is connected to the structured core by a short flexible loop and several side-chain interactions (Fig. 3b).

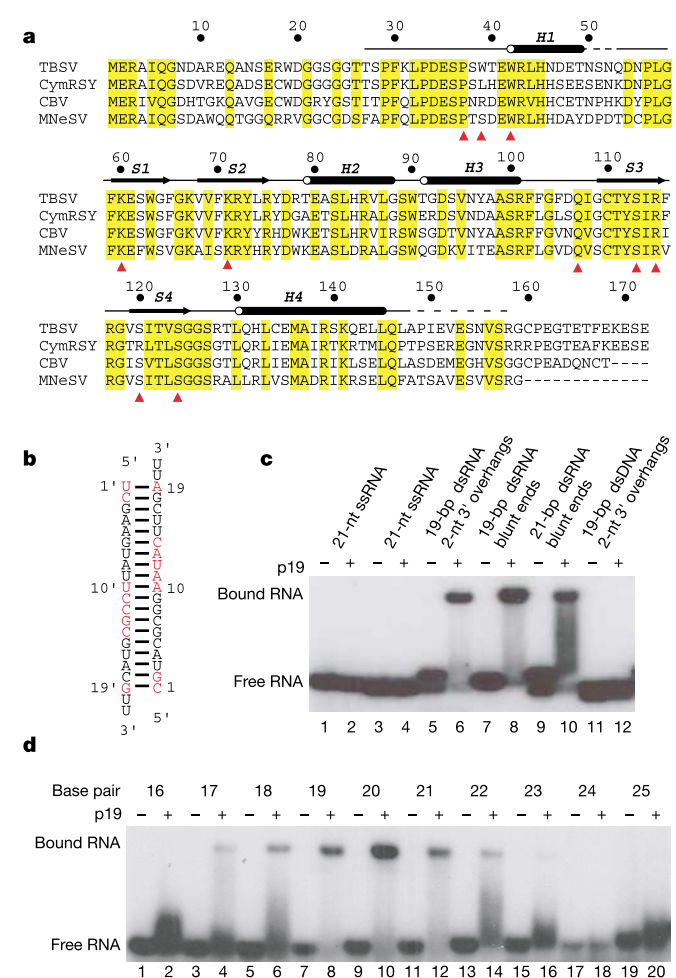


Figure 1 p19 and siRNA. **a**, Sequence alignment of p19 from the Tombusviridae family. The viruses with abbreviation and GenBank access numbers are as follows: tomato bushy stunt virus (TBSV) 979033; Cymbidium ringspot virus (CymRSV) 20087029; cucumber bulgarian virus (CBV) 30018257; maize necrotic streak virus (MNeSV) 10644292. Residues 27–158 from TBSV p19 were cloned and used in crystallization. Secondary structural elements in the X-ray structure are shown on the top. Dashed lines denote disordered residues. Conserved residues are coloured yellow. Residues directly contacting the siRNA duplex are marked by red triangles. **b**, The sequence of 21-nt siRNA used for crystallization. Residues contacting (direct and water-mediated) by the p19 homodimer are marked in red. **c**, Electrophoretic mobility shift assays for binding of p19 to RNA/DNA, with nucleic acid construct descriptions listed on the top. The bands in some lanes migrating faster than the duplex correspond to unannealed ssRNA. The majority of these ssRNAs form a duplex in the presence of p19 and are shifted to the bound RNA position. **d**, Electrophoretic mobility shift assays defining the length dependence of self-complementary siRNA duplexes for complex formation with p19. These self-complementary RNA sequences adopt either duplexes with 2-nt 3'-overhangs or alternate hairpin folds. The hairpin RNAs, which do not bind p19, are not shown.

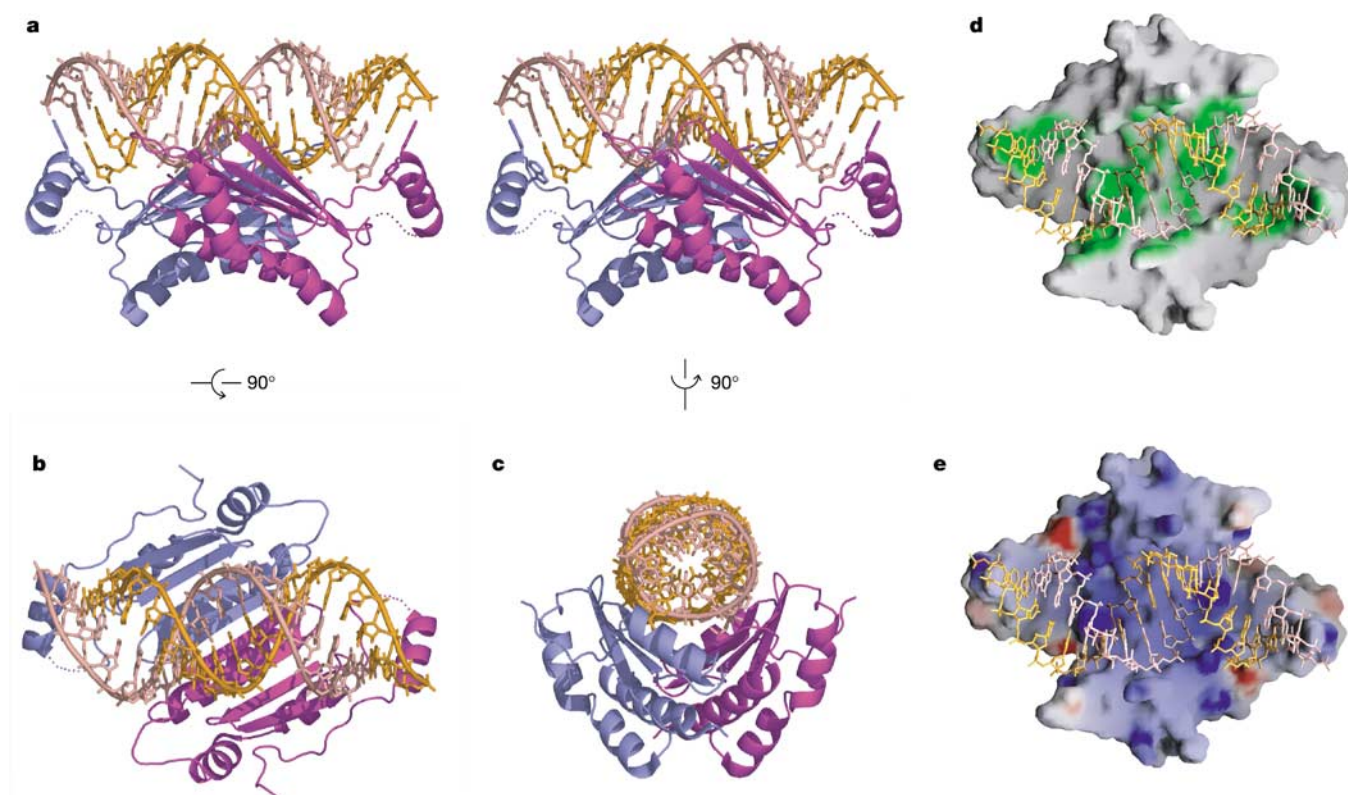


Figure 2 p19-siRNA complex structure. **a**, A stereo view perpendicular to the two-fold axis. Individual monomers of the p19 dimer are coloured blue and magenta, while the two siRNA strands are coloured orange and pink. Two tryptophans from each monomer of the p19 dimer, which bracket the terminal base pairs at either end of the siRNA duplex, are

shown in stick representation. **b, c**, Alternative mono views of the complex rotated by 90° along a different axis. **d**, RNA-contacting region (green) along the dimeric protein surface. **e**, Electrostatic surface of dimeric protein, with blue and red colours corresponding to positively and negatively charged patches.

The saddle-like β -sheet surface of the p19 homodimer spans one and a half turns of helix formed by the 21-nt (19-bp) siRNA duplex. Intermolecular contacts cover the central minor groove and two adjacent partial major grooves of the RNA duplex, with Fig. 4c showing four ribonucleotides from each RNA strand crossing four β -strands (S3'-S4'-S4-S3) in the p19 dimer, with a crossing angle of $\sim 40^\circ$ between the RNA backbone and protein β -strands.

Several direct and water-mediated intermolecular contacts involving non-bridging oxygens of the phosphodiester backbone have been observed in the structure of the complex. Three long side-chain residues (K71, R115 and Q107') form direct contacts with phosphate groups (11', 12'-13' and 14', respectively), while both backbone (CO of K67 and G118 and N of G126') and side-chain (S120) residues form water-mediated contacts with four phosphate groups (Fig. 4b).

The RNA minor groove-facing surface of the p19 central β -sheet is rich in serine and threonine (a total of ten per p19 dimer) residues (Fig. 4c). These hydroxyl-carrying residues plus four trapped water molecules (two OH groups per water molecule) form an unusual hydroxyl-group network that hydrogen-bonds with a total of six sugar 2'-OH groups lining the RNA minor groove. Serine and threonine (S113, T122, S120, S124' and T111') residues form direct and water-mediated intermolecular contacts with the 2'-OH groups on one RNA strand, with symmetry-related counterparts contacting the 2'-OH groups on the partner RNA strand (Fig. 4c). An additional 2'-OH group from each RNA strand is recognized by the backbone carbonyl groups of Q107/Q107' in the complex (Fig. 4c).

The p19 amino acids involved in direct recognition of the siRNA duplex are highlighted (red triangles) in Fig. 1a, and the majority

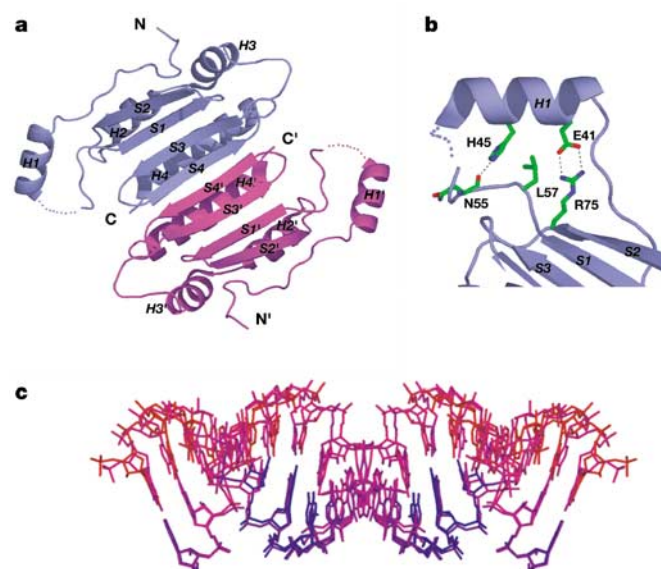


Figure 3 p19 and siRNA structures in complex. **a**, Ribbon representation of the p19 dimer in the complex. The β -strands are labelled S1 to S4 and the α -helices are labelled H1 to H4. The symmetry-related blue and magenta p19 monomers are designated by unprimed and primed symbols. The dimer forms a continuous eight-stranded β -sheet, which cradles the siRNA duplex within its concave face. The 'reading head' helices H1 and H1' extend from opposite edges of the β -sheet core. **b**, The side-chain interactions that position helix H1 relative to the β -sheet core of each monomer in the p19 dimer in the complex. **c**, siRNAs with opposite orientations in the complex, with colour-coding from blue to red, reflecting an increase in temperature factors. The RNA was oriented as in Fig. 2a.

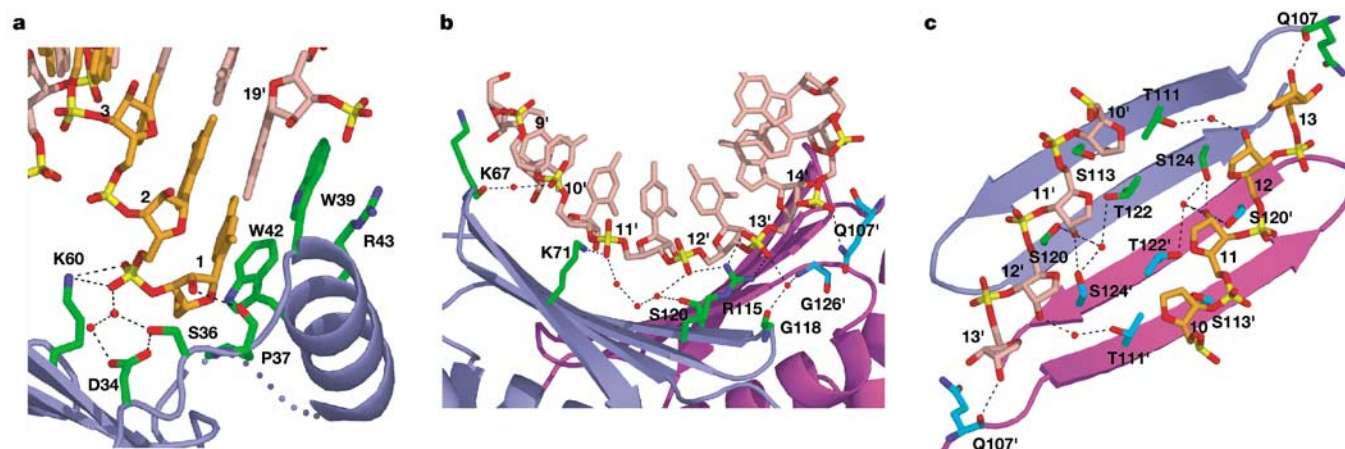


Figure 4 Details of p19-siRNA interactions. **a**, Stacking and hydrogen-bond/electrostatic recognition of RNA ends. **b**, Hydrogen-bonding/electrostatic recognition associated with the RNA major groove. Only one strand of the siRNA is shown for clarity. **c**, Hydrogen-bonding recognition associated with the RNA minor groove. Only the sugar-phosphate

backbone is shown for clarity. The interacting protein side chains and RNA sugar-phosphate backbones are coloured as follows: oxygens in red, nitrogens in navy blue, phosphates in yellow, carbons from protein monomer I in green, and carbons from monomer II in cyan. Other colour schemes are as Fig. 2.

of these are highly conserved amongst p19 homologues. The sequence conservation amongst p19 homologues suggests that other members of p19 family bind siRNA in a similar way. The siRNA sugar-phosphate backbone contacted by the p19 homodimer is highlighted (in red) in Fig. 1b, with the contacts distributed towards the centre and ends of the siRNA duplex. All electrostatic and hydrogen-bonding interactions are directed towards the sugar-phosphate backbone in the p19 dimer-siRNA complex, thereby readily explaining the sequence-independent nature of the recognition. The observed stacking of tryptophan pairs over the terminal base pairs also reflects sequence-independent contacts in the complex. Moreover, the critical contribution of 2'-OH group recognition in the complex correlates with the ability of p19 to distinguish between RNA and DNA duplexes.

The principles of protein architecture and molecular recognition observed for the p19 homodimer-siRNA complex are unique and distinct from previous structures of double-stranded RNA-binding domain (dsRBD) protein-RNA duplex complexes^{9,10}, where the proteins bound as monomers and loop and helical residues were involved in recognition of the sugar-phosphate backbone, and to a lesser extent of the base edges along the minor groove.

siRNAs, which are the products of dsRNA cleavage by the RNase III nuclease Dicer, play key roles in the RNA silencing pathway. siRNAs, which may form transient complexes with Dicer, are known to interact with proteins within the RNA-induced silencing complex (RISC)¹¹, in a process whereby one of the siRNA strands targets complementary mRNA. siRNAs may also serve as primers for the activity of host RNA-dependent RNA polymerases (RdRP), which have been implicated in the amplification and spreading of the silencing signal¹²⁻¹⁶. Because the molecular mechanisms underlying these interactions are currently unknown, it remains to be established whether the siRNA recognition principles defined from our studies of the p19 viral suppressor complex could govern recognition events in these other siRNA complexes. Nevertheless, our structure elegantly demonstrates how viral proteins can recognize and sequester siRNAs of defined length.

An improved understanding of viral suppression of RNA silencing in plants could have broad implications for manipulation of the RNA silencing machinery across species. Our structure of the p19 homodimer-siRNA complex provides molecular insights into siRNA-specific recognition processes and sets the stage for the rational engineering of p19 variants capable of siRNA targeting

with improved affinity, thereby enhancing available tools to manipulate RNA silencing events in gene regulation. □

Methods

Protein and RNA preparation

The p19 gene (1-172 residues) of TBSV was designed with codon usage optimized for expression in *Escherichia coli*, synthesized by a polymerase chain reaction (PCR)-based method¹⁷, and cloned into pET28a (Novagen) vector containing a thrombin-cleavable His-tag at the protein N terminus. A p19 fragment containing 27-158 residues and double methionine mutations at L144 and L147—which are introduced for facilitating phasing by multiple-wavelength anomalous diffraction (MAD)—referred to as p19m, was generated using PCR and the Quickchange site-directed mutagenesis kit (Stratagene). The truncation and mutations had no effect on the ability of p19 to bind the siRNA duplex. Selenomethionine (Se-Met)-substituted p19m was cultured using the *E. coli* strain BL21(DE3) in M9 minimal medium, as described previously¹⁸. Recombinant proteins were purified by a Ni-chelating affinity column, followed by His-tag removal with thrombin and additional purification by heparin chromatography. RNA oligonucleotides were chemically synthesized and subsequently deprotected, purified by denaturing polyacrylamide gel electrophoresis, and annealed to form siRNA duplexes. The stoichiometry of complex formation between p19m and siRNA was determined by mobility shift assays.

Crystallization and structure determination

Crystals were grown at 20 °C from hanging drops consisting of 1 µl each of complex (125 µM in 0.1 M KCl, 5 mM HEPES-KOH, 10 mM DTT, pH 7.6) and reservoir solution (1.2-1.6 M ammonium sulphate, 0.1 M HEPES-NaOH, pH 7.5). Crystals were harvested into 1.6 M ammonium sulphate, 0.1 M HEPES-NaOH, pH 7.5, 20% glycerol before flash freezing in liquid nitrogen. A three-wavelength MAD data set was collected on a Se-Met crystal on beamline 14IDB at the Advanced Photon Source (APS) at Argonne National laboratory. The crystals, of R32 space group ($a = b = 91.25$ Å, $c = 148.63$ Å), diffracted to 1.85 Å resolution. Integration, scaling and merging of the diffraction data were done with the HKL2000 suite of programs¹⁹. Four selenium sites and initial phases were determined using the CNS program²⁰. After density modification, the electron density map calculated to 1.95 Å was of excellent quality (Supplementary Fig. S1). 80% of protein residues could be automatically built into the map by the Resolve program²¹. The remaining protein and the RNA model was built using the program O²². The refinement was done using the CNS program against the data collected at the peak wavelength using a maximum-likelihood amplitude target. Overall anisotropic temperature factors and bulk solvent corrections were applied throughout the refinement.

The asymmetric unit contains one p19 monomer and half of the siRNA duplex. The 21-nt (19-bp) siRNA duplex (Fig. 1b) lacks two-fold symmetry. Therefore, the RNA was modelled as a 19-bp duplex with half occupancy, with a second duplex of opposite orientation generated by crystallographic symmetry. Refinement was done in space group R32 by explicitly turning off van der Waals packing interactions amongst the two oppositely oriented RNA duplexes. The sugar-phosphate backbones of oppositely oriented RNA duplexes were superimposable in the starting RNA model but separated during the structural refinement. The degree of separation correlates with the observed temperature factors, which are lowest (blue, Fig. 3c) for the protein-contacting segments, and highest (red, Fig. 3c) for the solvent-exposed segments. We anticipate that the associated separation and temperature factors are both indicative of mobility within the bound RNA.

The structure of the complex was refined to an R factor of 21.4% and $R_{\text{free}} = 23.6\%$ (for crystallographic statistics, see Supplementary Table S1). The final model includes protein residues 25–49 and 53–148, a total of 38 RNA nucleotides, 70 water molecules and 2 sulphate ions. N-terminal residues H25 and M26 are from the vector construct. Residues H25, D54, E80, F105, Q131 are modelled as partial side chains, whereas side chains of residues T111, S124, L133 and Q142 exhibit double conformations. Figures were prepared with PyMOL (<http://pymol.sourceforge.net/>) and GRASP²³, and the RNA duplex curvature was estimated using CURVES²⁴.

Electrophoretic mobility shift assays

The RNA duplexes (sequences available on request) were annealed as described⁶, and 5'-end-labelled with ³²P. The protein–RNA binding reactions contained 100 fmol RNA duplex, 10 nmol (as monomer) full-length p19 and 5 μ l 0.1 M KCl, 25 mM HEPES, 10 mM DTT, pH 7.6. After 15 min at room temperature, 1 μ l 50% glycerol and dye were added to the reaction products and separated in a 5% polyacrylamide gel in 25 mM Tris, 192 mM glycine, pH 8.3 at room temperature.

Received 28 September; accepted 14 November 2003; doi:10.1038/nature02213.

Published online 3 December 2003.

1. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature* **391**, 806–811 (1998).
2. Hannon, G. J. RNA interference. *Nature* **418**, 244–251 (2002).
3. Voinnet, O. RNA silencing as a plant immune system against viruses. *Trends Genet.* **17**, 449–459 (2001).
4. Li, W. X. & Ding, S. W. Viral suppressors of RNA silencing. *Curr. Opin. Biotechnol.* **12**, 150–154 (2001).
5. Voinnet, O., Pinto, Y. M. & Baulcombe, D. C. Suppression of gene silencing: a general strategy used by diverse DNA and RNA viruses of plants. *Proc. Natl Acad. Sci. USA* **96**, 14147–14152 (1999).
6. Silhavy, D. *et al.* A viral protein suppresses RNA silencing and binds silencing-generated, 21- to 25-nucleotide double-stranded RNAs. *EMBO J.* **21**, 3070–3080 (2002).
7. Bernstein, E., Caudy, A. A., Hammond, S. M. & Hannon, G. J. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**, 363–366 (2001).
8. Elbashir, S. M., Martinez, J., Patkaniowska, A., Lendeckel, W. & Tuschl, T. Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate. *EMBO J.* **20**, 6877–6888 (2001).
9. Rytter, J. M. & Schultz, S. C. Molecular basis of double-stranded RNA–protein interactions: structure of a dsRNA-binding domain complexed with dsRNA. *EMBO J.* **17**, 7505–7513 (1998).
10. Ramos, A. *et al.* RNA recognition by a Staufen double-stranded RNA-binding domain. *EMBO J.* **19**, 997–1009 (2000).
11. Hammond, S. M., Bernstein, E., Beach, D. & Hannon, G. J. An RNA-directed nuclease mediates post-transcriptional gene silencing in *Drosophila* cells. *Nature* **404**, 293–296 (2000).
12. Dalmay, T., Hamilton, A., Rudd, S., Angell, S. & Baulcombe, D. C. An RNA-dependent RNA polymerase gene in *Arabidopsis* is required for posttranscriptional gene silencing mediated by a transgene but not by a virus. *Cell* **101**, 543–553 (2000).
13. Mourrain, P. *et al.* *Arabidopsis* SGS2 and SGS3 genes are required for posttranscriptional gene silencing and natural virus resistance. *Cell* **101**, 533–542 (2000).
14. Lipardi, C., Wei, Q. & Paterson, B. M. RNAi as random degradative PCR: siRNA primers convert mRNA into dsRNAs that are degraded to generate new siRNAs. *Cell* **107**, 297–307 (2001).
15. Sijen, T. *et al.* On the role of RNA amplification in dsRNA-triggered gene silencing. *Cell* **107**, 465–476 (2001).
16. Himer, C., Dunoyer, P., Moissiard, G., Ritzenthaler, C. & Voinnet, O. Transitivity-dependent and -independent cell-to-cell movement of RNA silencing. *EMBO J.* **22**, 4523–4533 (2003).
17. Casimiro, D. R., Wright, P. E. & Dyson, H. J. PCR-based gene synthesis and protein NMR spectroscopy. *Structure* **5**, 1407–1412 (1997).
18. Doublié, S. Preparation of selenomethionyl proteins for phase determination. *Methods Enzymol.* **276**, 523–530 (1997).
19. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
20. Brünger, A. T. *et al.* Crystallography & NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54**, 905–921 (1998).
21. Terwilliger, T. C. Automated main-chain model-building by template-matching and iterative fragment extension. *Acta Crystallogr. D* **59**, 34–44 (2002).
22. Jones, T. & Kjeldgaard, M. Electron-density map interpretation. *Methods Enzymol.* **227**, 174–208 (1997).
23. Nicholls, A., Sharp, K. A. & Honig, B. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281–296 (1991).
24. Lavery, R. & Sklenar, H. The definition of generalized helicoidal parameters and of axis curvature for irregular nucleic acids. *J. Biomol. Struct. Dyn.* **6**, 63–91 (1988).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We thank members of the Patel laboratory for stimulating discussions and A. Teplov for assistance with data collection. This research was supported by NIH. We thank the personnel at beamline 14IDB of the Advanced Photon Source (APS) for their assistance. Use of this APS beamline was supported by the US Department of Energy, Basic Energy Sciences, Office of Science.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to D.P. (pateld@mshcc.org). Coordinates have been deposited in the Protein Data Bank under accession code 1R9F.

A conspicuous nickel protein in microbial mats that oxidize methane anaerobically

Martin Krüger¹, Anke Meyerdierks¹, Frank Oliver Glöckner¹, Rudolf Amann¹, Friedrich Widdel¹, Michael Kube², Richard Reinhardt², Jörg Kahnt³, Reinhard Böcher³, Rudolf K. Thauer³ & Seigo Shima³

¹Max Planck Institute for Marine Microbiology, Celsiusstraße 1, 28359 Bremen, Germany

²Max Planck Institute for Molecular Genetics, Ihnestraße, 14195 Berlin, Germany

³Max Planck Institute for Terrestrial Microbiology, Karl-von-Frisch-Straße, 35043 Marburg, Germany

Anaerobic oxidation of methane (AOM) in marine sediments is an important microbial process in the global carbon cycle and in control of greenhouse gas emission. The responsible organisms supposedly reverse the reactions of methanogenesis^{1–8}, but cultures providing biochemical proof of this have not been isolated. Here we searched for AOM-associated cell components in microbial mats from anoxic methane seeps in the Black Sea^{9–11}. These mats catalyse AOM rather than carry out methanogenesis. We extracted a prominent nickel compound displaying the same absorption spectrum as the nickel cofactor F₄₃₀ of methyl-coenzyme M reductase, the terminal enzyme of methanogenesis¹²; however, the nickel compound exhibited a higher molecular mass than F₄₃₀. The apparent variant of F₄₃₀ was part of an abundant protein that was purified from the mat and that consists of three different subunits. Determined amino-terminal amino acid sequences matched a gene locus cloned from the mat. Sequence analyses revealed similarities to methyl-coenzyme M reductase from methanogenic archaea. The abundance of the nickel protein (7% of extracted proteins) in the mat suggests an important role in AOM.

Large reservoirs of methane exceeding in mass conventional fossil fuels lie buried in deep, sulphate-depleted marine sediments^{13,14}. Despite permanent geochemical or microbial production and upward migration, little of the methane ever escapes into the free ocean water. In the upper subsurface sediment, where sulphate but no oxygen is present, most methane is scavenged by anaerobic microbial oxidation^{2,4,6,15–19} according to $\text{CH}_4 + \text{SO}_4^{2-} \rightarrow \text{HCO}_3^- + \text{HS}^- + \text{H}_2\text{O}$. Several *in situ* analyses based on isotope and lipid signatures^{5,11,13,20–22} or phylogenetic marker DNA or RNA^{5,7,10,11,14,22} have suggested that the microorganisms responsible for AOM represent special, hitherto uncultivated groups of archaea in association with sulphate-reducing bacteria. Most of these archaea belong to the ANME-1 and ANME-2 clusters. Both clusters are related to the Methanosarcinales^{5,7,11,14,22}. The associated, tentative sulphate-reducing bacteria usually affiliate with the *Desulfosarcina/Desulfococcus* cluster of the Deltaproteobacteria.

Besides unique microorganisms, AOM involves intriguing mechanisms. There is still a dispute about how methane oxidation is linked to sulphate reduction^{2,6,23,24}. Furthermore, the free energy gain from AOM (ΔG *in situ* estimated between -10 and -40 kJ mol⁻¹) is one of the smallest known to fuel the metabolism of a microbial life-form^{2,21,23,24}. Moreover, methane is chemically unreactive. Mechanisms for its biochemical activation are, therefore, of particular interest.

A natural system that seems to be well suited for a cultivation-independent biochemical study of AOM is microbial mats from the cold anoxic methane seeps of the northwestern Black Sea shelf^{9–11}. These mats provide sufficient biomass to record a cytochrome spectrum⁹. They harbour archaea of the ANME-1 cluster that account for up to 70% of the cells detectable *in situ*¹¹. Cells of the

Desulfosarcina cluster account for 25%, with the remaining proportion accounted for by other microorganisms. During incubation in anoxic sea water at *in situ* temperature (12 °C), these mats catalysed methane-dependent sulphate reduction at a specific rate of approximately 0.1 nmol min⁻¹ (mg protein)⁻¹. Methanogenesis measured without added methane or with conventional methanogenic substrates (H₂ + CO₂, acetate, methanol, formate, or methylamines) was only approximately 10% of this rate. This indicates that the mats are dedicated to AOM rather than to methanogenesis.

In the present study, we focused on cell components with a putative involvement in the catalysis of the initial step of AOM. Our investigations were based on the hypothesis¹⁻⁸ that this step is essentially a reversal of the terminal reaction in methanogenesis; that is, the reduction of methyl-coenzyme M (CoM-S-CH₃) with coenzyme B (H-S-CoB), yielding methane (CH₄) and the heterodisulphide (CoM-S-S-CoB)¹². The reaction is catalysed by methyl-coenzyme M reductase (MCR), a 300 kDa enzyme composed of three different subunits (α₂β₂γ₂) and two tightly but non-covalently bound molecules of a nickel porphyrinoid, cofactor F₄₃₀, with a molecular mass of 905 Da (Supplementary Fig. 1). The assumption of reversed methanogenesis through such an enzyme (or a similar one) is supported by a recent analysis of genes retrieved from habitats with AOM that harbour ANME-1 and ANME-2 populations⁸; the deduced amino acid sequences were highly similar to those of the three subunits of MCR. The involvement of a similar enzyme in AOM is also indicated by our finding that methane-dependent sulphate reduction by the mats was inhibited by bromoethanesulphonate (>1 mM), a specific inhibitor of MCR.

Chromatography of mat extracts yielded a fraction with the same absorption spectrum (ultraviolet/visible) as authentic cofactor F₄₃₀ (Supplementary Fig. 2). However, mass spectrometry showed that this fraction contained two nickel compounds at a ratio of 1:4 (peak

intensities) and at a molecular mass of 905 and 951 Da, respectively (Fig. 1a). The presence of nickel was indicated by the relative peak intensities of the isotopomers, which were as predicted from the natural Ni isotope distribution (68.3% ⁵⁸Ni, 26.1% ⁶⁰Ni, 1.1% ⁶¹Ni, 3.6% ⁶²Ni, 0.9% ⁶⁴Ni). The mass spectrum of the lighter nickel compound was identical to that of authentic F₄₃₀, whereas the mass spectrum of the heavier one did not match that of a known cofactor. We subsequently searched for the 951 Da nickel compound in sediments from two other habitats where AOM takes place: Hydrate Ridge (northeast Pacific) and Eckernförde Bay (Baltic Sea). These also contained the 951 Da nickel compound in addition to the 905 Da compound. As a control, 18 species of methanogenic archaea representing major phylogenetic clusters in addition to sludge from a methanogenic sewage digester were analysed. In these, only the 905 Da nickel compound was detected.

Around 5 nmol of the 951 Da nickel compound was extracted per gram of wet microbial mat. Any protein that harbours this compound should be present in the mat in noticeable amounts. Indeed, after protein pre-purification by ammonium sulphate fractionation, ion exchange chromatography revealed a prominent protein peak (Fig. 1b) that contained the 951 Da nickel compound (Supplementary Fig. 3a) and was, as with MCR, composed of three different subunits (Fig. 1b, inset). For convenience and for differentiation from a much less abundant protein containing the 905 Da nickel compound (Supplementary Fig. 3b), the prominent protein is referred to as Ni-protein I. Approximately 0.4 mg Ni-protein I was obtained per gram of wet mat. Assuming that the composition of this protein is analogous to that of known MCR, the extracted amount of Ni-protein I (1.33 nmol) is expected to contain 2.66 nmol of the 951 Da nickel compound. The deviation from the totally extracted amount (see above) may be explained by losses due to protein adsorption by cell debris (including exopolymers) and partial precipitation during ammonium sulphate fractionation. The proportion of Ni-protein I was calculated to amount to approximately 7% of the extracted protein, assuming that all main cell types in the mat were lysed before protein extraction (see Methods).

Edman degradation of the N-terminal parts of the three subunits showed that the purified Ni-protein I was homogeneous. The N-terminal sequences almost perfectly matched those deduced

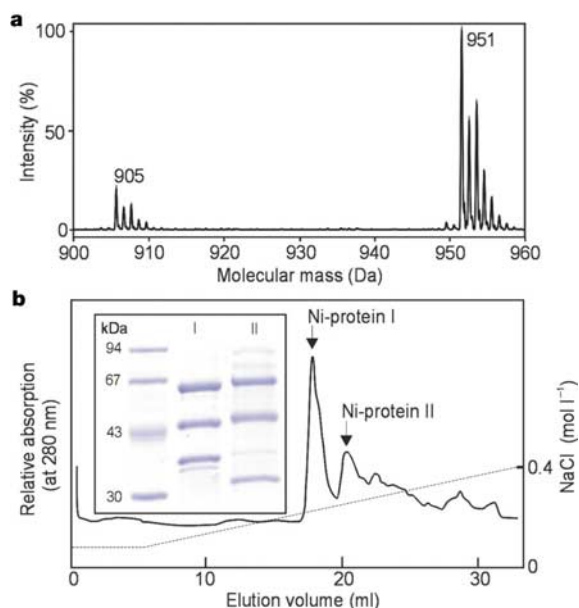


Figure 1 Biochemical analysis of a microbial mat that oxidizes methane anaerobically. **a**, MALDI-TOF mass spectrum of a yellowish fraction purified from the mat. The spectrum reveals two compounds with main masses at 905 and 951 Da, and isotopomer intensities indicative of nickel. The 905 Da compound resembles the nickel cofactor F₄₃₀ from methanogenic archaea. **b**, Anion exchange chromatography of native mat proteins yielding two distinct proteins, Ni-protein I and II, containing the 951 and 905 Da compound, respectively. Denaturing gel electrophoresis (inset, with molecular mass markers) reveals three different subunits in Ni-proteins I and II.

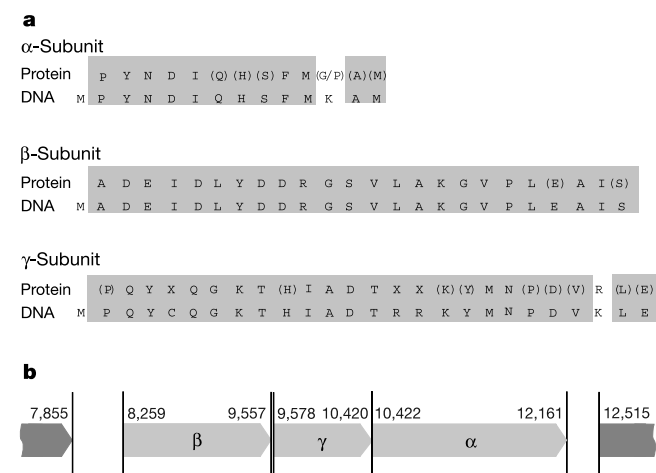


Figure 2 Identification of the genes encoding the dominant protein, Ni-protein I, in the mat with the capacity for anaerobic oxidation of methane. **a**, Comparison of N-terminal amino acid sequences determined chemically in the three subunits of Ni-protein I and amino acid sequences deduced from genes in a DNA fragment from the mat. Uncertain positions are in parentheses. **b**, Order of the juxtaposed genes that encode the three subunits of Ni-protein I (see also Supplementary Fig. 4).

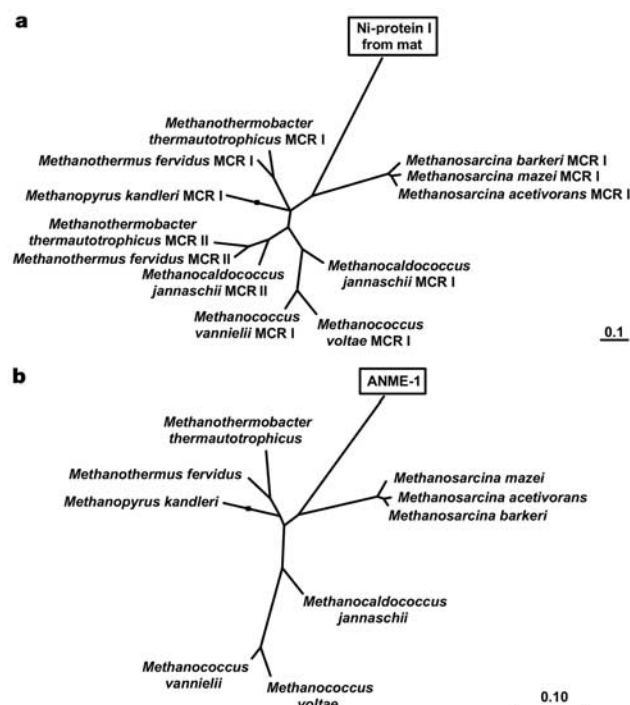


Figure 3 Phylogenetic relationships between Ni-protein I and MCR, and between the ANME-1 cluster and methanogenic archaea. See Supplementary Fig. 6 for details. **a**, Consensus tree created from the trees of the three subunits of Ni-protein I from the mat and authentic MCR type I and II from various methanogenic archaea (see also Supplementary Fig. 6). The multifurcation point could not be unambiguously resolved on the basis of the available data set. **b**, Corresponding 16S rRNA tree of the respective archaea.

from three open reading frames on a 34.8-kilobase (kb) contig retrieved from the microbial mat (Fig. 2a). The encoding genes were part of an apparent transcription unit (Fig. 2b). The deduced amino acid sequences of the three subunits of Ni-protein I revealed a high degree of similarity with those of methyl-coenzyme M reductases from various methanogenic archaea (Supplementary Fig. 5). Nevertheless, Ni-protein I presents a distinct line of descent from an evolutionary point of view (Fig. 3a; see also Supplementary Fig. 6). With respect to gene organization and deduced amino acid sequences, the genes encoding Ni-protein I are most similar to those recently retrieved from other habitats with AOM and ANME-1 populations⁸.

Amino acid alignments between Ni-protein I and MCR subunits, with consideration of the three-dimensional structure of the latter¹², show that the active-site amino acids on the β - and γ -subunits are conserved. In the α -subunit, however, Ni-protein I differs from MCR by a motif (VX₂CCX₄CX₅C) with a valine (underlined) instead of the active-site glutamine, and three cysteine residues (underlined) adjacent to but not within the assumed active site.

The contig with the identified genes of the subunits of Ni-protein I and a contig (66.4 kb) with the analysed 16S ribosomal RNA gene and adjacent reading frames of the ANME-1 cluster showed similar numerical DNA characteristics. Correspondence analysis of synonymous codon usage²⁵ revealed clear relatedness of both contigs; the plot of relative codon frequencies yielded a correlation coefficient of 0.93 between both. The G + C content of the contigs was 42.9 and 43.0 mol%, respectively. Moreover, tetranucleotide frequency analysis^{26,27} showed a correlation coefficient of 0.73 between the two contigs.

Our data provide several arguments for the assumption that

Ni-protein I is associated with the ANME-1 cells and has a catalytic role in AOM. First, Ni-protein I accounted for as much as 7% of extractable proteins. Even in the case of a pure culture, a protein with such a proportion would be regarded as very abundant; it must be regarded as even more abundant if attributed to a cell type in a mixed population. Contribution of this proportion by the major cell type ANME-1 is thus a much more plausible explanation than contribution by the less abundant cell types. Second, a primary role of Ni-protein I in methanogenesis, even though not ruled out, is unlikely because a 1,000-fold lower amount or less would be fully sufficient for the observed rates of methane formation (as concluded from activities of known MCR¹²). Third, the similar codon usage, G+C content and tetranucleotide frequencies of the above contigs also indicate that Ni-protein I is associated with ANME-1. Fourth, phylogenetic trees show a marked coincidence between the branching of Ni-protein I among known MCR types (Fig. 3a; see also Supplementary Fig. 6) and the branching of 16S rRNA from ANME-1 among 16S rRNAs of methanogenic archaea (Fig. 3b). Furthermore, the recent genetic analysis of other habitats with AOM led to deduced protein sequences with distinct branching among MCR types⁸. One may speculate that during evolution of the ANME-1 archaea a parallel evolution of Ni-protein I occurred towards an efficient catalysis of AOM. In conclusion, with the available data, Ni-protein I is a likely candidate for the enzyme that catalyses methane activation as the crucial step in AOM (Supplementary Fig. 7). Ultimate proof of this assumption needs the isolation and biochemical study of cultures from the mats. □

Methods

Sources of biological material

Mat material was collected during the cruise of the project GHOSTDABS (University of Hamburg) in July 2001. The mats had a water content of approximately 90%, a nitrogen content of approximately 0.8% and a CaCO₃ content of approximately 4%. Methanogenic archaea subjected to cofactor analyses were *Methanosarcina acetivorans*, *M. barkeri*, *M. mazei*, *M. thermophila*, *Methanoseta concillii* and *Methanohalobium evestigatum* (Methanosarcinales); *Methanobacterium bryantii*, *Methanobrevibacter arboripophilus*, *Methanothermobacter marburgensis*, *M. thermoautotrophicus*, *Methanothermobacter fervidus* and *Methanospirillum stadtmanae* (Methanobacteriales); *Methanococcus thermophilus*, *Methanogenium cariaci* and *Methanospirillum hungatei* (Methanomicrobiales); *Methanocaldococcus voltae* and *Methanocaldococcus jannaschii* (Methanococcales); and *Methanopyrus kandleri* (Methanopyrales). Strains were obtained from the German Collection of Microorganisms and Cell Cultures (DSMZ), Braunschweig. Anaerobic sewage sludge was from the municipal sewage plant of Marburg.

Purification of nickel compounds

All purification steps were performed at 4 °C. Approximately 10 g (wet mass) of microbial mat were suspended in 10 ml of 50 mM MOPS/KOH, pH 7.0, ultrasonicated (30 min, 80 W) under cooling, and centrifuged (10 min, 4,100g). The pellet was resuspended in 20 ml H₂O, acidified to pH 3 with formic acid, and ultrasonicated and centrifuged (as before). The supernatants from both centrifugations were combined and centrifuged again (45 min, 150,000g). The new supernatant was adjusted to pH 7.5 with NaOH and applied to a QAE 25 Sepharose column (1 cm × 5 cm) that had been equilibrated with 50 mM Tris/HCl, pH 7.5. The column was washed with 120 ml H₂O and eluted with 5 mM formic acid while light absorption of the eluent was monitored at 430 nm. The light-absorbing (yellowish) fraction was loaded on a PAD-1 reversed-phase column (0.5 cm × 5 cm) that had been equilibrated with 50 mM formic acid/NaOH, pH 3. The column was washed with 27 mM formic acid and eluted with 60% (vol./vol.) methanol in water. The light-absorbing fraction was concentrated by lyophilization. Nickel-containing compounds were quantified photometrically using the molar absorption coefficient of authentic F₄₃₀ (23,000 cm⁻¹ M⁻¹ at 430 nm). Molecular masses were determined by matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry.

Protein purification

All purification steps were performed at 4 °C. A total of 32 g microbial mat (wet mass) was suspended in 60 ml of 50 mM MOPS/KOH, pH 7.0, decanted from settled rock particles (4.3 g), ultrasonicated (20 min, 80 W) and centrifuged (30 min, 40,000g). Sixty millilitres of the supernatant (with approximately 180 mg protein) were slowly supplemented with dry ammonium sulphate to 80% saturation. Precipitated proteins (approximately 160 mg) were removed by centrifugation (20 min, 10,000g). Samples of 10 ml (approximately 3 mg protein) were desalted and applied on a Uno Q1 column (35 mm × 7 mm), which was eluted with an NaCl gradient (0–0.4 M) in 50 mM MOPS/KOH, pH 7.0. Approximately 70% (2 mg) of the applied protein eluted in the first peak, which was Ni-protein I. Protein was quantified in the Coomassie-blue-binding assay

(versus bovine serum albumin as standard). Protein subunits were separated by sodium dodecylsulphate polyacrylamide gel electrophoresis.

Library construction

Genomic DNA was directly extracted²⁸ from the microbial mat and further purified by anion exchange chromatography. Fosmid libraries were constructed using the EpiFOS Fosmid Library Production Kit or the CopyControl Fosmid Library Production Kit (Epicentre). The libraries were screened for the gene encoding the α -subunit of MCR by colony hybridization using DIG-labelled polymerase chain reaction products and published primer sets²⁹.

DNA sequencing and annotation

Fosmids were sequenced by a shotgun approach based on plasmid libraries with 1.5- and 3.5-kb inserts; sequence coverage was more than tenfold. Resulting reads were assembled by Phrap44 and manually finished in GAP4 (<http://www.sanger.ac.uk>). The quality of the sequence data was finished to reach a maximum of one error in 10,000 bases. Finished fosmid sequences were automatically analysed with the software package GenDB³⁰ and manually annotated.

Phylogenetic reconstruction

Only full-length sequences were used for tree reconstruction. Protein tree topologies were tested by performing maximum likelihood, distance matrix and parsimony analyses in combination with different filter sets using the software ARB (<http://www.arb-home.de>). The 16S rRNA tree was reconstructed with ARB_Parsimony based on full-length sequences (total of 894 with >1,000 base pairs) and optimized by applying a 50% conservation filter. Partial sequences (total of 801) were inserted without allowing a change to the overall tree topology. For clarity, only sequences corresponding to the protein tree are shown.

Received 25 May; accepted 28 October 2003; doi:10.1038/nature02207.

- Zehnder, A. J. B. & Brock, T. D. Methane formation and methane oxidation by methanogenic bacteria. *J. Bacteriol.* **137**, 420–432 (1979).
- Hoehler, T. M., Alperin, M. J., Albert, D. B. & Martens, C. S. Field and laboratory studies of methane oxidation in an anoxic marine sediment: evidence for a methanogen-sulfate reducer consortium. *Global Biogeochem. Cycles* **8**, 451–463 (1994).
- Harder, J. Anaerobic methane oxidation by bacteria employing ¹⁴C-methane uncontaminated with ¹⁴C-carbon monoxide. *Mar. Geol.* **137**, 13–23 (1997).
- Hansen, L. B., Finster, K., Fossing, H. & Iversen, N. Anaerobic methane oxidation in sulfate depleted sediments: effects of sulfate and molybdate additions. *Aquat. Microbial Ecol.* **14**, 195–204 (1998).
- Hinrichs, K. U., Hayes, J. M., Sylva, S. P., Brewer, P. G. & DeLong, E. F. Methane-consuming archaeobacteria in marine sediments. *Nature* **398**, 802–805 (1999).
- Valentine, D. L. & Reeburgh, W. S. New perspectives on anaerobic methane oxidation. *Environ. Microbiol.* **2**, 477–484 (2000).
- Boetius, A. *et al.* A marine microbial consortium apparently mediating anaerobic oxidation of methane. *Nature* **407**, 623–626 (2000).
- Hallam, S. J., Girguis, P. G., Preston, C. M., Richardson, P. M. & DeLong, E. Identification of methyl coenzyme M reductase A (*mcrA*) genes associated with methane-oxidizing archaea. *Appl. Environ. Microbiol.* **69**, 5483–5491 (2003).
- Pimenov, N. V. *et al.* Bacterial mats on coral-like structures at methane seeps in the Black Sea. *Mikrobiologiya (Moscow)* **66**, 354–360 (1997).
- Tourova, T. P., Kolganova, T. V., Kuznetsov, B. B. & Pimenov, N. V. Phylogenetic diversity of the archaeal component in microbial mats on coral-like structures associated with methane seeps in the Black Sea. *Mikrobiologiya (Moscow)* **71**, 196–201 (2002).
- Michaelis, W. *et al.* Microbial reefs in the Black Sea fueled by anaerobic oxidation of methane. *Science* **297**, 1013–1015 (2002).
- Thauer, R. K. Biochemistry of methanogenesis: a tribute to Marjory Stephenson. *Microbiology* **144**, 2377–2406 (1998).
- Elvert, M. & Suess, E. Anaerobic methane oxidation associated with marine gas hydrates: Superlight C-isotopes from saturated and unsaturated C₂₀ and C₂₅ irregular isoprenoids. *Naturwissenschaften* **86**, 295–300 (1999).
- Orphan, V. J., House, C. H., Hinrichs, K. U., McKeegan, K. D. & DeLong, E. F. Multiple archaeal groups mediate methane oxidation in anoxic cold seep sediments. *Proc. Natl Acad. Sci. USA* **99**, 7663–7668 (2002).
- Martens, C. S. & Berner, R. A. Methane production in the interstitial waters of sulfate-depleted marine sediments. *Science* **185**, 1167–1169 (1974).
- Barnes, R. & Goldberg, E. Methane production and consumption in anoxic marine sediments. *Geology* **4**, 297–300 (1976).
- Reeburgh, W. Methane consumption in Cariaco Trench waters and sediments. *Earth Planet. Sci. Lett.* **28**, 345–352 (1976).
- Iversen, N. & Jørgensen, B. B. Anaerobic methane oxidation rates at the sulfate-methane transition in marine sediments from Kattegat and Skagerrak (Denmark). *Limnol. Oceanogr.* **30**, 944–955 (1985).
- Niewohner, C., Hensen, C., Kasten, S., Zabel, M. & Schulz, H. D. Deep sulfate reduction completely mediated by anaerobic methane oxidation in sediments of the upwelling area off Namibia. *Geochim. Cosmochim. Acta* **62**, 455–464 (1998).
- Pancost, R. D., Hopmans, E. C. & Damste, J. S. S. Archaeal lipids in Mediterranean cold seeps: Molecular proxies for anaerobic methane oxidation. *Geochim. Cosmochim. Acta* **65**, 1611–1627 (2001).
- Bian, L. *et al.* Algal and archaeal polyisoprenoids in a recent marine sediment: molecular-isotopic evidence for anaerobic oxidation of methane. *Geochim. Geophys. Geosyst.* **2**, 1–22 (2001).
- Orphan, V. J., House, C. H., Hinrichs, K. U., McKeegan, K. D. & DeLong, E. F. Methane-consuming archaea revealed by directly coupled isotopic and phylogenetic analysis. *Science* **293**, 484–487 (2001).
- Nauhaus, K., Boetius, A., Krüger, M. & Widdel, F. *In vitro* demonstration of anaerobic oxidation of methane coupled to sulphate reduction in sediment from a marine gas hydrate area. *Environ. Microbiol.* **4**, 296–305 (2002).
- Sørensen, K. B., Finster, K. & Ramsing, N. B. Thermodynamic and kinetic requirements in anaerobic methane oxidizing consortia exclude hydrogen, acetate, and methanol as possible electron shuttles. *Microbial Ecol.* **42**, 1–10 (2001).
- Lynn, D. J., Singer, G. A. & Hickey, D. A. Synonymous codon usage is subject to selection in thermophilic bacteria. *Nucleic Acids Res.* **30**, 4272–4277 (2002).
- Pride, D. T., Meinersmann, R. J., Wassenaar, T. M. & Blaser, M. J. Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res.* **13**, 145–158 (2003).
- Schbath, S. An efficient statistic to detect over- and under-represented words in DNA sequences. *J. Comput. Biol.* **4**, 189–192 (1997).
- Zhou, J., Bruns, M. A. & Tiedje, J. M. DNA recovery from soils of diverse composition. *Appl. Environ. Microbiol.* **62**, 316–322 (1996).
- Lueders, T., Chin, K.-J., Conrad, R. & Friedrich, M. Molecular analyses of methyl-coenzyme M reductase alpha-subunit (*mcrA*) genes in rice field soil and enrichment cultures reveal the methanogenic phenotype of a novel archaeal lineage. *Environ. Microbiol.* **3**, 194–204 (2001).
- Meyer, F. *et al.* GenDB—an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res.* **31**, 2187–2195 (2003).

Supplementary Information accompanies the paper on www.nature.com/nature.

Acknowledgements We are particularly grateful to W. Michaelis, K. Nauhaus, R. Seifert and the Professor Logachev shipboard party for providing microbial mat samples, and to A. Boetius for coordination work and advice. We thank J. Moll, R. Appel and M. Sordel-Klippert for technical assistance; J. Knecht for the chemical analyses; D. Linder for N-terminal sequencing; and H. Teeling and T. Lombardot for help with bioinformatics. This work is part of the projects MUMM, GHOSTDABS and the GenoMic network Göttingen, supported by the Federal Ministry of Education and Research (Germany). Further support came from the Max Planck Society. This is a publication of GHOSTDABS and of the programme GEOTECHNOLOGIEN of the Federal Ministry of Education and Research and the Deutsche Forschungsgemeinschaft.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to F.W. (fviddel@mpi-bremen.de) or R.K.T. (thauer@mail.uni-marburg.de). The nucleotide sequence of the contig from the Black Sea mat has been deposited at EMBL GenBank under accession number BX649197.

nature insight

protein misfolding

nature
insight
protein misfolding



Cover illustration

Representations of the structural ensembles of the folding transition states for a group of small proteins. (Courtesy of K. Lindorff-Larsen and M. Vendruscolo.)

At school it all sounded so simple — transcription turns DNA into RNA, and translation of RNA gives you protein. But the often forgotten third step in this process, the folding of the translated linear strand of amino acids into a fully functional three-dimensional protein, is one of the most complex challenges facing the cellular protein factory.

Although it has long been known that the amino-acid sequence in some way dictates the biologically active conformation of a protein, the experimental tools required to probe the intermediate states along the folding pathway have only begun to become available in the past decade or so. These tools are revealing a tightly regulated assembly line, where multiple factors guide nascent proteins to select the correct shape from an almost infinite array of possibilities.

Becoming apparent are the stringent quality-control systems that come into play if the folding process fails, ensuring that the misfolded products are targeted for degradation before they cause harm. Those that escape this cellular surveillance are prone to forming aggregates that can damage or kill cells through mechanisms that are just beginning to be understood.

A huge variety of previously unrelated diseases, such as prion diseases, diabetes and cancer, share the pathological feature of aggregated misfolded protein deposits. This suggests the exciting possibility that these 'protein-misfolding diseases' are linked by common principles, and may therefore present common targets for therapeutic intervention.

The articles in this Insight give an interdisciplinary overview of the field of protein-folding research, treading a path from protein chemistry through cell biology to misfolding diseases and the potential for therapeutic development.

This Insight is entirely self-contained; however, its scope and focus were mirrored in the organization of last year's Horizon Symposium on 'Protein Folding and Disease', the first in a new series of discussion meetings created as a joint initiative by Nature Publishing Group and Aventis. Complementary material can therefore be found at www.horizonsymposia.com.

We are pleased to acknowledge the financial support of Aventis in producing this Insight. As always, *Nature* carries sole responsibility for all editorial content and peer review.

Adam Smith Editor, *Nature Reviews Drug Discovery*

review article:

884 Protein folding and misfolding

C. M. Dobson

progress:

891 Quality control in the endoplasmic reticulum protein factory

R. Sitia & I. Braakman

review article:

895 Protein degradation and protection against misfolded or damaged proteins

A. L. Goldberg

review article:

900 Folding proteins in fatal ways

D. J. Selkoe

review article:

905 Therapeutic approaches to protein-misfolding diseases

F. E. Cohen & J. W. Kelly

Editor, *Nature*: Philip Campbell
Publisher: Peter Collins
Insights Editor: Lesley Anson
Production Editor: Maria Hodges

Art Editor: Martin Harrison
Layouts: Nadia Rooney
Production: Sue Gray
Sponsorship: Mark Greene

Diagrams: Charlotte Cartwright
Mick Ayres
Nik Spencer

Protein folding and misfolding

Christopher M. Dobson

University of Cambridge, Department of Chemistry, Lensfield Road, Cambridge CB2 1EW, UK (e-mail: cmd44@cam.ac.uk)

The manner in which a newly synthesized chain of amino acids transforms itself into a perfectly folded protein depends both on the intrinsic properties of the amino-acid sequence and on multiple contributing influences from the crowded cellular milieu. Folding and unfolding are crucial ways of regulating biological activity and targeting proteins to different cellular locations. Aggregation of misfolded proteins that escape the cellular quality-control mechanisms is a common feature of a wide range of highly debilitating and increasingly prevalent diseases.

One of the defining characteristics of a living system is the ability of even the most intricate of its component molecular structures to self-assemble with precision and fidelity. Uncovering the mechanisms through which such processes take place is one of the grand challenges of modern science¹. The folding of proteins into their compact three-dimensional structures is the most fundamental and universal example of biological self-assembly; understanding this complex process will therefore provide a unique insight into the way in which evolutionary selection has influenced the properties of a molecular system for functional advantage. The wide variety of highly specific structures that result from protein folding and that bring key functional groups into close proximity has enabled living systems to develop astonishing diversity and selectivity in their underlying chemical processes. In addition to generating biological activity, however, we now know that folding is coupled to many other biological processes, including the trafficking of molecules to specific cellular locations and the regulation of cellular growth and differentiation². In addition, only correctly folded proteins have long-term stability in crowded biological environments and are able to interact selectively with their natural partners. It is therefore not surprising that the failure of proteins to fold correctly, or to remain correctly folded, is the origin of a wide variety of pathological conditions. In this article we explore the underlying mechanism of protein folding and of the nature and consequences of misfolding and its links with disease.

The fundamental mechanism of protein folding

The concept of an energy landscape

The mechanism by which a polypeptide chain folds to a specific three-dimensional protein structure has until recently been shrouded in mystery. Native states of proteins almost always correspond to the structures that are most thermodynamically stable under physiological conditions³. Nevertheless, the total number of possible conformations of any polypeptide chain is so large that a systematic search for this particular structure would take an astronomical length of time. However, it is now clear that the folding process does not involve a series of mandatory steps between specific partly folded states, but rather a stochastic search of the many conformations accessible to a polypeptide chain^{3–5}. The inherent fluctuations in the conformation of an unfolded or incompletely folded polypeptide chain enable even residues that are highly separated in the amino-acid sequence to come into contact with one other. Because, on average, native-like interactions between residues are more stable than non-native ones, they are more persistent and the polypeptide chain is able to find its lowest-energy structure by a process of

trial and error. Moreover, if the energy surface or 'landscape' has the right shape (see Fig. 1) only a small number of all possible conformations needs to be sampled by any given protein molecule during its transition from a random coil to a native structure^{3–6}. Because the landscape is encoded by the amino-acid sequence, natural selection has enabled proteins to evolve so that they are able to fold rapidly and efficiently.

Such a description, based more on the ideas of statistical mechanics and polymer physics than on those of classic chemical dynamics, is often referred to as the 'new view' of protein folding⁷. As well as providing a firm conceptual basis for folding, it has shown that many of the earlier phenomenological descriptions of the folding process are important limiting cases of a general mechanism. These ideas are stimulating the investigation of the most elementary steps in the folding process by both experimental and theoretical procedures. For example, biophysical measurements and computer simulations have revealed that many of the local elements of protein structures can be generated very rapidly; for example, individual α -helices are able to form in less than 100 ns, and β -turns in as little as 1 μ s (refs 8, 9). Indeed, the folding *in vitro* of some of the simplest proteins, such as small helical bundles, is completed in less than 50 μ s (refs 10, 11). Intriguingly, some other small proteins, particularly those based on β -sheet structures, can take many orders of magnitude longer to fold, as we see below, but such rate changes can be understood to a significant extent in terms of the characteristics of the native structures¹².

A key question is how does the correct fold emerge from such fundamental steps; that is, how is the energy landscape unique to a specific protein defined by its amino-acid sequence. The structural transitions taking place during folding *in vitro* can be investigated in detail by a variety of techniques, ranging from optical methods to NMR spectroscopy³, some of which can now even be used to follow the behaviour of single molecules¹³. The latter capability is of particular significance in the context of probing the stochastic nature of the folding process (see Fig. 1). Studies of a series of small proteins, typically with 60–100 residues, have been crucial for investigating the most basic steps in folding because these proteins convert from their unfolded states to their native states without the complication of highly populated intermediates. For these systems, monitoring the effects of specific mutations on the kinetics of folding and unfolding has proved to be a seminal technique, because of its ability to probe the role of individual residues in the folding process¹⁴. Particular insight has come from the use of this approach to analyse the transition states for folding, namely the critical regions of energy surfaces through which all molecules must pass to reach the native fold (see Fig. 1). The results of many studies of these species suggest that the

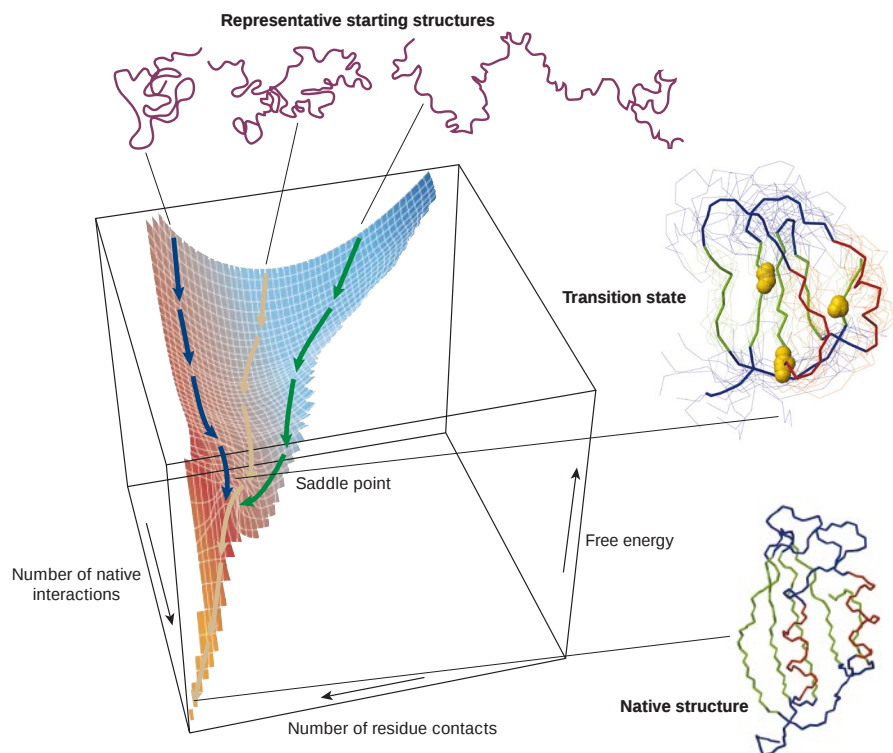


Figure 1 A schematic energy landscape for protein folding. The surface is derived from a computer simulation of the folding of a highly simplified model of a small protein. The surface 'funnels' the multitude of denatured conformations to the unique native structure. The critical region on a simple surface such as this one is the saddle point corresponding to the transition state, the barrier that all molecules must cross if they are to fold to the native state. Superimposed on this schematic surface are ensembles of structures corresponding to different stages of the folding process. The transition state ensemble was calculated by using computer simulations constrained by experimental data from mutational studies of acylphosphatase¹⁸. The yellow spheres in this ensemble represent the three 'key residues' in the structure; when these residues have formed their native-like contacts the overall topology of the native fold is established. The structure of the native state is shown at the bottom of the surface; at the top are indicated schematically some contributors to the distribution of unfolded species that represent the starting point for folding. Also indicated on the surface are highly simplified trajectories for the folding of individual molecules. Adapted from ref. 6.

fundamental mechanism of protein folding involves the interaction of a relatively small number of residues to form a folding nucleus, about which the remainder of the structure rapidly condenses¹⁵.

More details of how such a mechanism is able to generate a unique fold have emerged from a range of theoretical studies, particularly involving computer simulation techniques¹⁶. Of particular significance are investigations that compare the simulation results with experimental observations^{6,17}. One approach incorporates experimental measurements directly into the simulations as restraints limiting the regions of conformational space that are explored in each simulation; this strategy has enabled rather detailed structures to be generated for transition states¹⁸ (see Fig. 1). The results suggest that, despite a high degree of disorder, these structures have the same overall topology as the native fold. In essence, interactions involving the key residues force the chain to adopt a rudimentary native-like architecture. Although it is not yet clear exactly how the sequence encodes such characteristics, the essential elements of the fold are likely to be determined primarily by the pattern of hydrophobic and polar residues that favours preferential interactions of specific residues as the structure becomes increasingly compact. Once the correct topology has been achieved, the native structure will then almost invariably be generated during the final stages of folding¹⁸. Conversely, if these key interactions are not formed, the protein cannot fold to a stable globular structure; this mechanism therefore acts also as a 'quality-control' process by which misfolding can generally be avoided.

The determinants of protein folds

Secondary structure, the helices and sheets that are found in nearly every native protein structure, is stabilized primarily by hydrogen

bonding between the amide and carbonyl groups of the main chain. The formation of such structure is an important element in the overall folding process, although it might not have as fundamental a role as the establishment of the overall chain topology¹⁹. Perhaps the most dramatic evidence for such a conclusion is the observation of a remarkable correlation between the experimental folding rates of a wide range of small proteins and the complexity of their folds, measured by the contact order¹². The latter is the average separation in the sequence between residues that are in contact with each other in the native structure. The existence of such a correlation can be rationalized by the argument that a stochastic search process will be more time consuming if the residues that form the nucleus are further away from each other in the sequence. This evidence strongly supports the conclusion that there are relatively simple underlying principles by which the sequence of a protein encodes its structure²⁰. Not only will the establishment of such principles reveal in more depth how proteins are able to fold, but it should advance significantly our ability to predict protein folds directly from their sequences and to design sequences that encode novel folds.

For proteins with more than about 100 residues, experiments generally reveal that one (or more) intermediate is significantly populated during the folding process. There has, however, been considerable discussion about the significance of such species: whether they assist the protein to find its correct structure or whether they are traps that inhibit the folding process^{21–23}. Regardless of the outcome of this debate, the structural properties of intermediates provide important evidence about the folding of these larger proteins. In particular, they suggest that these proteins generally fold in modules, in other words, folding can take place largely independently in different segments or domains of the protein^{6,14}. In such cases, interactions involving key

residues are likely to establish the native-like fold within local regions or domains and also to ensure that the latter then interact appropriately to form the correct overall structure^{23,24}. The fully native structure is only acquired when all the native-like interactions have been formed both within and between the domains; this happens in a final cooperative folding step when all the side chains become locked in their unique close-packed arrangement and water is excluded from the protein core²⁵. This modular mechanism is appealing because it suggests that highly complex structures might be assembled in manageable pieces. Moreover, such a principle can readily be extended to describe the assembly of other macromolecules, particularly nucleic acids, and even large 'molecular machines' such as the ribosome.

Protein folding and misfolding in the cell

In a cell, proteins are synthesized on ribosomes from the genetic information encoded in the cellular DNA. Folding *in vivo* is in some cases co-translational; that is, it is initiated before the completion of protein synthesis, whereas the nascent chain is still attached to the ribosome²⁶. Other proteins, however, undergo the major part of their folding in the cytoplasm after release from the ribosome, whereas yet others fold in specific compartments, such as mitochondria or the endoplasmic reticulum (ER), after trafficking and translocation through membranes^{27,28}. Many details of the folding process depend on the particular environment in which folding takes place, although the fundamental principles of folding, discussed above, are undoubtedly universal. But because incompletely folded proteins must inevitably expose to the solvent at least some regions of structure that are buried in the native state, they are prone to inappropriate interaction with other molecules within the crowded environment of a cell²⁹. Living systems have therefore evolved a range of strategies to prevent such behaviour^{27–29}.

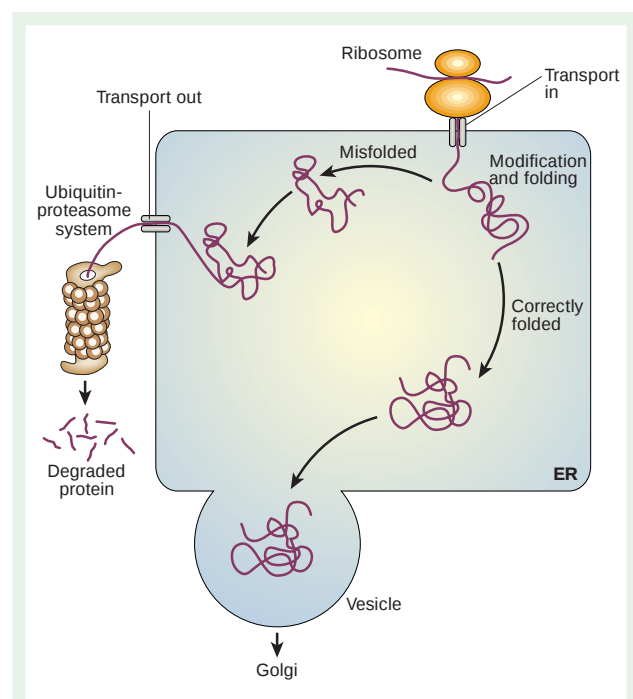


Figure 2 Regulation of protein folding in the ER. Many newly synthesized proteins are translocated into the ER, where they fold into their three-dimensional structures with the help of a series of molecular chaperones and folding catalysts (not shown). Correctly folded proteins are then transported to the Golgi complex and then delivered to the extracellular environment. However, incorrectly folded proteins are detected by a quality-control mechanism and sent along another pathway (the unfolded protein response) in which they are ubiquitinated and then degraded in the cytoplasm by proteasomes. Adapted from ref. 32.

Of particular importance in this context are the many molecular chaperones that are present in all types of cells and cellular compartments. Some chaperones interact with nascent chains as they emerge from the ribosome, whereas others are involved in guiding later stages of the folding process^{27,28}. Molecular chaperones often work in tandem to ensure that the various stages in the folding of such systems are all completed efficiently. Many of the details of the functions of molecular chaperones have been determined from studies of their effects on folding *in vitro*. The best characterized of the chaperones studied in this manner is the bacterial complex involving GroEL, a member of the family of 'chaperonins', and its 'co-chaperone' GroES. Many aspects of the sophisticated mechanism through which this coupled system functions are now well understood^{27,28}. Of particular interest is that GroEL, and other members of this class of molecular chaperone, contains a cavity in which incompletely folded polypeptide chains can enter and undergo the final steps in the formation of their native structures while sequestered and protected from the outside world.

Molecular chaperones do not themselves increase the rate of individual steps in protein folding; rather, they increase the efficiency of the overall process by reducing the probability of competing reactions, particularly aggregation. However, there are several classes of folding catalyst that accelerate potentially slow steps in the folding process. The most important are peptidylprolyl isomerases, which increase the rate of *cis-trans* isomerization of peptide bonds involving proline residues, and protein disulphide isomerases, which enhance the rate of formation and reorganization of disulphide bonds³⁰. Despite these factors, given the enormous complexity and the stochastic nature of the folding process, it would be remarkable if misfolding never occurred. Clear evidence that molecular chaperones are needed to prevent misfolding and its consequences comes from the fact that the concentrations of many of these species are substantially increased during cellular stress; indeed, the designation of many as heat shock proteins (Hsps) reflects this fact. It is also clear that some molecular chaperones are able not only to protect proteins as they fold but also to rescue misfolded and even aggregated proteins and enable them to have a second chance to fold correctly^{27,28}. Active intervention in the folding process requires energy, and ATP is required for most of the molecular chaperones to function with full efficiency.

In eukaryotic systems, many of the proteins that are synthesized in a cell are destined for secretion to the extracellular environment. These proteins are translocated into the ER, where folding takes place before secretion through the Golgi apparatus. The ER contains a wide range of molecular chaperones and folding catalysts, and in addition the proteins that fold here must satisfy a 'quality-control' check before being exported (Fig. 2)^{31,32}. Such a process is particularly important because there seem to be few molecular chaperones outside the cell, although one (clusterin), at least, has recently been discovered³³. This quality-control mechanism involves a remarkable series of glycosylation and deglycosylation reactions that enables correctly folded proteins to be distinguished from misfolded ones³¹. The importance of these regulatory systems is underlined by recent experiments that suggest that a large fraction of all polypeptide chains synthesized in a cell fail to pass this test and are targeted for degradation³⁴. Like the 'heat shock response' in the cytoplasm, the 'unfolded protein response' in the ER is also stimulated (upregulated) during stress and, as we shall see below, is strongly linked to the avoidance of misfolding diseases³⁵.

Folding and unfolding are the ultimate ways of generating and abolishing specific types of cellular activity. In addition, processes as apparently diverse as translocation across membranes, trafficking, secretion, the immune response and regulation of the cell cycle are directly dependent on folding and unfolding events². Failure to fold correctly, or to remain correctly folded, will therefore give rise to the malfunctioning of living systems and hence to disease^{36–38}. Some of these diseases (such as cystic fibrosis³⁶ and some types of cancer³⁹) result from proteins folding incorrectly and not being able to exercise

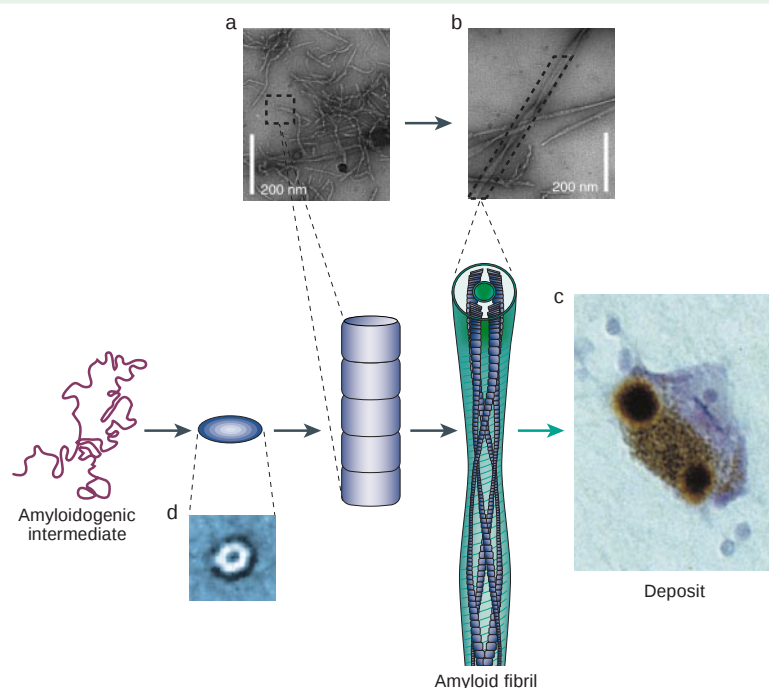


Figure 3 A schematic representation of the general mechanism of aggregation to form amyloid fibrils. Unfolded or partially unfolded proteins associate with each other to form small, soluble aggregates that undergo further assembly into protofibrils or protofilaments (**a**) and then mature fibrils (**b**, top electron microscope images from ref. 44). The fibrils often accumulate in plaques or other structures such as the Lewy bodies associated with Parkinson's disease (**c**, electron microscope image on right). Some of the early aggregates seem to be amorphous or micellar in nature, although others form ring-shaped species with diameters of approximately 10 nm (**d**, electron microscope image from ref. 53). Adapted from a figure provided by H. A. Lashuel and P. T. Lansbury Jr.

their proper function; many such disorders are familial because the probability of misfolding is often greater in mutational variants. In other cases, proteins with a high propensity to misfold escape all the protective mechanisms and form intractable aggregates within cells or (more commonly) in extracellular space. An increasing number of disorders, including Alzheimer's and Parkinson's diseases, the spongiform encephalopathies and type II diabetes, are directly associated with the deposition of such aggregates in tissues, including the brain, heart and spleen^{37,38,40,41}. In the next section we look at the formation of these species.

Protein aggregation and amyloid formation

Each amyloid disease involves predominantly the aggregation of a specific protein, although a range of other components including additional proteins and carbohydrates are incorporated into the deposits when they form *in vivo*. In neurodegenerative diseases, the quantities of aggregates involved can sometimes be so small as to be almost undetectable, whereas in some systemic diseases literally kilograms of protein can be found in one or more organs⁴⁰. The characteristics of the soluble forms of the 20 or so proteins involved in the well-defined amyloidoses are very varied — they range from intact globular proteins to largely unstructured peptide molecules — but the aggregated forms have many characteristics in common⁴². Amyloid deposits all show specific optical behaviour (such as birefringence) on binding certain dye molecules such as Congo red. The fibrillar structures typical of many of the aggregates have very similar morphologies (long, unbranched and often twisted structures a few nanometres in diameter) and a characteristic 'cross- β ' X-ray fibre diffraction pattern. The latter reveals that the organized core structure is composed of β -sheets whose strands run perpendicular to the fibril axis⁴². The ability of polypeptide chains to form amyloid structures is not restricted to the relatively small number of proteins associated with recognized clinical disorders, and it now seems to be a generic feature of polypeptide chains^{37,43}. The most compelling evidence for the latter statement is that fibrils can be formed *in vitro* by many other peptides and proteins, including such well-known molecules as myoglobin, and also by homopolymers such as polythreonine or polylysine^{37,44}.

Although no structure of an amyloid fibril has yet been determined in atomic detail, increasingly convincing models based on data from techniques such as X-ray fibre diffraction⁴², cryoelectron microscopy⁴⁵ and solid-state NMR⁴⁶ are emerging. The core structure of the fibrils seems to be stabilized primarily by interactions, particularly hydrogen bonds, involving the polypeptide main chain. Because the main chain is common to all polypeptides, this observation explains why fibrils formed from polypeptides of very different sequence seem to be so similar^{42,43}. In some cases only a handful of the residues of a given protein might be involved in this structure, with the remainder of the chain being associated in some other manner with the fibrillar assembly; in other cases almost the whole polypeptide chain seems to be involved. The generic amyloid structure contrasts strongly with the highly individualistic globular structures of most natural proteins. In these latter structures the interactions associated with the very specific packing of the side chains seem to override the main-chain preferences^{43,44}.

Even though the ability to form amyloid fibrils seems to be generic, the propensity to do so under given circumstances can vary markedly between different sequences. The relative aggregation rates for a wide range of peptides and proteins correlates with the physicochemical features of the molecules such as charge, secondary-structure propensities and hydrophobicity⁴⁷. In a globular protein the polypeptide main chain and the hydrophobic side chains are largely buried within the folded structure. Only when they are exposed, for example when the protein is partly unfolded (for example, at low pH) or fragmented (for example, by proteolysis), will conversion into amyloid fibrils be possible. Experiments *in vitro* indicate that their formation is then generally characterized by a lag phase, followed by a period of rapid growth^{48,49}. Such behaviour is typical of nucleated processes such as crystallization; the lag phase can be eliminated by the addition of preformed aggregates to fresh solutions, a process known as seeding. An interesting recent suggestion is that seeding by chemically modified forms of proteins, resulting for example from deamidation or oxidative stress, might in some cases be an important factor in triggering the aggregation process and the onset of disease⁵⁰.

There are striking similarities in the aggregation behaviour of different peptides and proteins (Fig. 3)^{48,49}. The first phase in amyloid

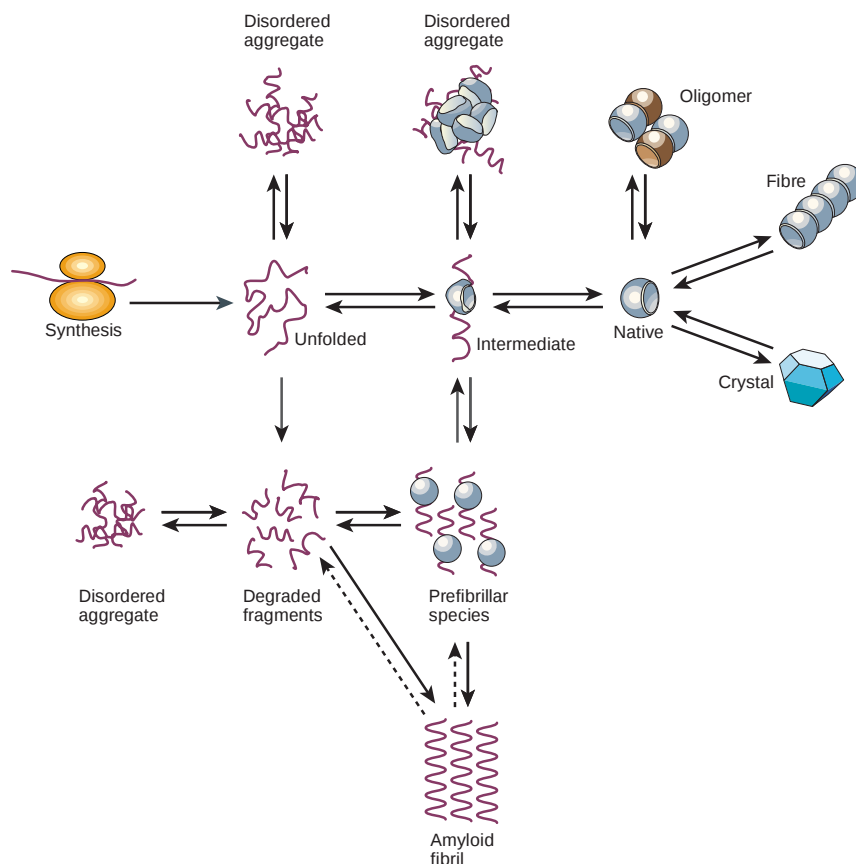


Figure 4 A unified view of some of the types of structure that can be formed by polypeptide chains. An unstructured chain, for example newly synthesized on a ribosome, can fold to a monomeric native structure, often through one or more partly folded intermediates. It can, however, experience other fates such as degradation or aggregation. An amyloid fibril is just one form of aggregate, but it is unique in having a highly organized 'misfolded' structure, as shown in Fig. 3. Other assemblies, including functional oligomers, macromolecular complexes and natural protein fibres, contain natively folded molecules, as do the protein crystals produced *in vitro* for X-ray diffraction studies of their structures. The populations and interconversions of the various states are determined by their relative thermodynamic and kinetic stabilities under any given conditions. In living systems, however, transitions between the different states are highly regulated by the environment and by the presence of molecular chaperones, proteolytic enzymes and other factors. Failure of such regulatory mechanisms is likely to be a major factor in the onset and development of misfolding diseases. Adapted from ref. 54.

formation seems to involve the formation of soluble oligomers as a result of relatively nonspecific interactions, although, in some cases, specific structural transitions, such as domain swapping, might be important⁵¹. The earliest species visible by electron or atomic-force microscopy generally resemble small bead-like structures, sometimes linked together, and often described as amorphous aggregates or as micelles. These early 'prefibrillar aggregates' then transform into species with more distinctive morphologies, often called 'protofilaments' or 'protofibrils'. These structures are commonly short, thin, sometimes curly, fibrillar species that are thought to assemble into mature fibrils, perhaps by lateral association accompanied by some degree of structural reorganization. The aggregates that form first are likely to be relatively disorganized structures that expose to the outside world a variety of segments of the protein that are normally buried in the globular state⁵². In some cases, however, these early aggregates appear to adopt quite distinctive structures, including well-defined annular species⁵³ (see Fig. 3).

Molecular evolution and the control of protein misfolding

The state of a protein that is adopted under specific conditions depends on the relative thermodynamic stabilities of the various accessible conformations and on the kinetics of their interconversion (Fig. 4)^{37,54}. Amyloid fibrils are just one of the types of aggregate that can be formed by proteins, although a significant feature of this particular species is that its highly organized hydrogen-bonded structure is likely to give it unique kinetic stability. Thus, once formed, such aggregates can persist for long periods, allowing a progressive build-up of deposits in tissue, and indeed enabling seeding of the subsequent conversion of additional quantities of the same protein into amyloid fibrils. It is therefore not surprising that biological sys-

tems have almost universally avoided the deliberate formation of such material. Nevertheless, there is increasing evidence that the unique properties of amyloid structures have been exploited by some species, including bacteria, fungi and even mammals, for specific (and carefully regulated) purposes^{55–57}.

There is evidence that evolutionary selection has tended to avoid amino-acid sequences, such as alternating polar and hydrophobic residues, that favour a β -sheet structure of the type seen in amyloid fibrils⁵⁸. Moreover, recent studies suggest that the aggregation process that results in amyloid fibrils is nucleated in a similar manner to that of folding, but that the residues involved might well be located in different regions of the sequence from those that nucleate folding⁵⁹. Such 'kinetic partitioning' means that mutations that occur during evolution could be selected for their ability to enhance folding at the expense of aggregation. However, it is apparent that biological systems have become robust not just by careful manipulation of the sequences of proteins but also by controlling, by means of molecular chaperones and degradation mechanisms, the particular state adopted by a given polypeptide chain at a given time and under given conditions. This process can be thought of as being analogous to the way in which biology regulates and controls the various chemical transformations that take place in the cell by means of enzymes. And just as the aberrant behaviour of enzymes can cause metabolic disease, the aberrant behaviour of the chaperone and other machinery regulating polypeptide conformations can contribute to misfolding and aggregation diseases^{35,60}.

The ideas encapsulated in Fig. 4 therefore serve as a framework for understanding the fundamental events that underlie misfolding diseases. For example, many of the mutations associated with the familial forms of deposition diseases increase the population of partially

unfolded states, and hence the propensity to aggregate, by decreasing the stability or cooperativity of the native state^{37,41,61,62}. Other familial diseases are associated with the accumulation of amyloid deposits whose primary components are fragments of native proteins; such fragments can be produced by aberrant processing or incomplete proteolysis, and are unable to fold into aggregation-resistant states. Other pathogenic mutations enhance the propensities of such species to aggregate, for example by increasing their hydrophobicity or decreasing their charge⁴⁷. And, in the prion disorders such as Kuru or Creutzfeldt–Jakob disease, it seems that ingestion of pre-aggregated states of an identical protein, for example by voluntary or involuntary cannibalism or through the use of contaminated pharmaceuticals or surgical instruments, can markedly increase the inherent rate of aggregation through seeding and hence can generate a mechanism for transmission^{48,63}.

In some aggregation diseases, the large quantities of insoluble protein involved can physically disrupt specific organs and thereby cause pathological behaviour⁴⁰. But for neurodegenerative disorders, such as Alzheimer's disease, the primary symptoms almost certainly result from a 'toxic gain of function' associated with aggregation⁶⁴. The early prefibrillar aggregates of proteins associated with such diseases are highly damaging to cells; by contrast, the mature fibrils are usually relatively benign^{48,65}. Moreover, experiments have recently suggested that similar aggregates of proteins that are not connected with any known diseases could be equally cytotoxic⁵². The generic nature of such aggregates and their effects on cells has recently been supported by the remarkable finding that antibodies can cross-react with early aggregates of different peptides and proteins, and moreover inhibit their toxicity⁶⁶. It is possible that there are specific mechanisms for this toxicity, for example as a result of annular species (Fig. 3) that resemble the toxins produced by bacteria that form pores in membranes and disrupt the ion balance in cells⁵³. However, it is likely that the relatively disorganized prefibrillar aggregates are also harmful to cells, probably through a less specific mechanism, for example as a result of the exposure of non-native hydrophobic surfaces stimulating aberrant interactions with cell membranes or other cellular components⁶⁷.

Future directions

In normal circumstances the molecular chaperones and other 'housekeeping' mechanisms are remarkably efficient in ensuring that such potentially toxic species as prefibrillar aggregates are neutralized before they can do any damage^{28,68}. This neutralization could result simply from the efficient targeting of misfolded proteins for degradation, but it seems that molecular chaperones are also able to alter the partitioning between harmful and harmless forms of aggregates (Fig. 4)⁶⁹. If the efficiency of these protective mechanisms is impaired, however, the probability of pathogenic behaviour increases^{35,68}. Such a process would explain why most of the amyloid diseases are associated with old age, when there is likely to be an increased tendency for proteins to become misfolded or damaged, coupled with a decreased efficiency of the molecular chaperone and unfolded proteins responses⁷⁰. It is ironic that through our success in increasing the life expectancy of the populations of the developed world, we are now seeing the limitations of our proteins and of the regulatory mechanisms that control their behaviour⁷¹. It is therefore essential that we use our developing understanding of misfolding and aggregation to find effective strategies for combating these increasingly common and highly debilitating diseases⁵⁴. Fortunately, there is now real evidence to suggest that modern science will rise successfully to this tremendous challenge. □

doi:10.1038/nature02261

- Vendruscolo, M., Zurdo, J., MacPhee, C. E. & Dobson, C. M. Protein folding and misfolding: a paradigm of self-assembly and regulation in complex biological systems. *Phil. Trans. R. Soc. Lond.* **361**, 1205–1222 (2003).
- Radford, S. E. & Dobson, C. M. From computer simulations to human disease: emerging themes in protein folding. *Curr. Opin. Struct. Biol.* **9**, 291–298 (1999).
- Dobson, C. M., Sali, A. & Karplus, M. Protein folding: a perspective from theory and experiment.

- Angew. Chem. Int. Ed. Eng.* **37**, 868–893 (1998).
- Wolynes, P. G., Onuchic, J. N. & Thirumalai, D. Navigating the folding routes. *Science* **267**, 1619–1620 (1995).
- Dill, K. A. & Chan, H. S. From Levinthal to pathways to funnels. *Nature Struct. Biol.* **4**, 10–19 (1997).
- Dinner, A. R., Sali, A., Smith, L. J., Dobson, C. M. & Karplus, M. Understanding protein folding via free energy surfaces from theory and experiment. *Trends Biochem. Sci.* **25**, 331–339 (2000).
- Baldwin, R. L. Protein folding: matching speed and stability. *Nature* **369**, 183–184 (1994).
- Eaton, W. A., Munoz, V., Thompson, P. A., Henry, E. R. & Hofrichter, J. Kinetics and dynamics of loops, α -helices, β -hairpins, and fast-folding proteins. *Acc. Chem. Res.* **31**, 745–753 (1998).
- Snow, C. D., Nguyen, H., Pande, V. S. & Gruebele, M. Absolute comparison of simulated and experimental protein-folding dynamics. *Nature* **420**, 102–106 (2002).
- Yang, W. Y. & Gruebele, M. Folding at the speed limit. *Nature* **423**, 193–197 (2003).
- Mayor, U. *et al.* The complete folding pathway of a protein from nanoseconds to microseconds. *Nature* **421**, 863–867 (2003).
- Plaxco, K. W., Simons, K. T. & Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985–994 (1998).
- Schuler, B., Lipman, E. A. & Eaton, W. A. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature* **419**, 743–747 (2002).
- Fersht, A. R. *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (W. H. Freeman, New York, 1999).
- Fersht, A. R. Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism. *Proc. Natl Acad. Sci. USA* **97**, 1525–1529 (2000).
- Shea, J. E. & Brooks, C. L. From folding surfaces to folding proteins: a review and assessment of simulation studies of protein folding and unfolding. *Annu. Rev. Phys. Chem.* **52**, 499–535 (2001).
- Fersht, A. R. & Daggett, V. Protein folding and unfolding at atomic resolution. *Cell* **108**, 573–582 (2002).
- Vendruscolo, M., Paci, E., Dobson, C. M. & Karplus, M. Three key residues form a critical contact network in a transition state for protein folding. *Nature* **409**, 641–646 (2001).
- Makarov, D. E. & Plaxco, K. W. The topomer search model: a simple, quantitative theory of two-state protein folding kinetics. *Protein Sci.* **12**, 17–26 (2003).
- Baker, D. A surprising simplicity to protein folding. *Nature* **405**, 39–42 (2000).
- Roder, H. & Colon, W. Kinetic role of early intermediates in protein folding. *Curr. Opin. Struct. Biol.* **7**, 15–28 (1997).
- Sanchez, I. E. & Kiefhaber, T. Evidence for sequential barriers and obligatory intermediates in apparent two-state protein folding. *J. Mol. Biol.* **325**, 367–376 (2003).
- Khan, F., Chuang, J. I., Gianni, S. & Fersht, A. R. The kinetic pathway of folding of barnase. *J. Mol. Biol.* **333**, 169–186 (2003).
- Vendruscolo, M., Paci, E., Karplus, M. & Dobson, C. M. Structures and relative free energies of partially folded states of proteins. *Proc. Natl Acad. Sci. USA* **100**, 14817–14821 (2003).
- Cheung, M. S., Garcia, A. E. & Onuchic, J. N. Protein folding mediated by solvation: water expulsion and formation of the hydrophobic core occur after the structural collapse. *Proc. Natl Acad. Sci. USA* **99**, 685–690 (2002).
- Hardesty, B. & Kramer, G. Folding of a nascent peptide on the ribosome. *Prog. Nucleic Acid Res. Mol. Biol.* **66**, 41–66 (2001).
- Bukau, B. & Horwich, A. L. The Hsp70 and Hsp60 chaperone machines. *Cell* **92**, 351–366 (1998).
- Hartl, F. U. & Hayer-Hartl, M. Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science* **295**, 1852–1858 (2002).
- Ellis, R. J. Macromolecular crowding: an important but neglected aspect of the intracellular environment. *Curr. Opin. Struct. Biol.* **11**, 114–119 (2001).
- Schiene, C. & Fischer, G. Enzymes that catalyse the restructuring of proteins. *Curr. Opin. Struct. Biol.* **10**, 40–45 (2000).
- Hammon, C. & Helenius, A. Quality control in the secretory pathway. *Curr. Opin. Cell Biol.* **7**, 523–529 (1995).
- Kaufman, R. J. *et al.* The unfolded protein response in nutrient sensing and differentiation. *Nature Rev. Mol. Cell Biol.* **3**, 411–421 (2002).
- Wilson, M. R. & Easterbrook Smith, S. B. Clusterin is a secreted mammalian chaperone. *Trends Biochem. Sci.* **25**, 95–98 (2000).
- Schubert, U. *et al.* Rapid degradation of a large fraction of newly synthesized proteins by proteasomes. *Nature* **404**, 770–774 (2000).
- Bence, N. F., Sampat, R. M. & Kopito, R. R. Impairment of the ubiquitin–proteasome system by protein aggregation. *Science* **292**, 1552–1555 (2001).
- Thomas, P. J., Qu, B. H. & Pedersen, P. L. Defective protein folding as a basis of human disease. *Trends Biochem. Sci.* **20**, 456–459 (1995).
- Dobson, C. M. The structural basis of protein folding and its links with human disease. *Phil. Trans. R. Soc. Lond. B* **356**, 133–145 (2001).
- Horwich, A. Protein aggregation in disease: a role for folding intermediates forming specific multimeric interactions. *J. Clin. Invest.* **110**, 1221–1232 (2002).
- Bullock, A. N. & Fersht, A. R. Rescuing the functions of mutant p53. *Nature Rev. Cancer* **1**, 68–76 (2001).
- Tan, S. Y. & Pepys, M. B. Amyloidosis. *Histopathology* **25**, 403–414 (1994).
- Kelly, J. W. Alternative conformation of amyloidogenic proteins and their multi-step assembly pathways. *Curr. Opin. Struct. Biol.* **8**, 101–106 (1998).
- Sunde, M. & Blake, C. C. F. The structure of amyloid fibrils by electron microscopy and X-ray diffraction. *Adv. Protein Chem.* **50**, 123–159 (1997).
- Dobson, C. M. Protein misfolding, evolution and disease. *Trends Biochem. Sci.* **24**, 329–332 (1999).
- Fändrich, M. & Dobson, C. M. The behaviour of polyamino acids reveals an inverse side-chain effect in amyloid structure formation. *EMBO J.* **21**, 5682–5690 (2002).
- Jiménez, J. L. *et al.* Cryo-electron microscopy of an SH3 amyloid fibril and model of the molecular packing. *EMBO J.* **18**, 815–821 (1999).
- Petkova, A. T. *et al.* A structural model for Alzheimer's β -amyloid fibrils based on experimental constraints from solid state NMR. *Proc. Natl Acad. Sci. USA* **99**, 16742–16747 (2002).
- Chiti, F., Stefani, M., Taddei, N., Ramponi, G. & Dobson, C. M. Rationalization of mutational effects on protein aggregation rates. *Nature* **424**, 805–808 (2003).
- Caughey, B. & Lansbury, P. T. Jr. Protofibrils, pores, fibrils, and neurodegeneration: separating the

- responsible protein aggregates from the innocent bystanders. *Annu. Rev. Neurosci.* **26**, 267–298 (2003).
49. Bitan, G. *et al.* Amyloid β -protein (A β) assembly: A β 40 and A β 42 oligomerize through distinct pathways. *Proc. Natl Acad. Sci. USA* **100**, 330–335 (2003).
50. Nilsson, M. R., Driscoll, M. & Raleigh, D. P. Low levels of asparagine deamidation can have a dramatic effect on aggregation of amyloidogenic peptides: implications for the study of amyloid formation. *Protein Sci.* **11**, 342–349 (2002).
51. Schlunegger, M. P., Bennett, M. J. & Eisenberg, D. Oligomer formation by 3D domain swapping: a model for protein assembly and misassembly. *Adv. Protein Chem.* **50**, 61–122 (1997).
52. Bucciantini, M. *et al.* Inherent cytotoxicity of aggregates implies a common origin for protein misfolding diseases. *Nature* **416**, 507–511 (2002).
53. Lashuel, H. A., Hartley, D., Petre, B. M., Walz, T. & Lansbury, P. T. Jr. Neurodegenerative disease: amyloid pores from pathogenic mutations. *Nature* **418**, 291 (2002).
54. Dobson, C. M. Protein folding and disease: a view from the First Horizon Symposium. *Nature Rev. Drug Discov.* **2**, 154–160 (2003).
55. True, H. L. & Lindquist, S. L. A yeast prion provides a mechanism for genetic variation and phenotypic diversity. *Nature* **407**, 477–483 (2000).
56. Chapman, M. R. *et al.* Role of *Escherichia coli* curli operons in directing amyloid fiber formation. *Science* **295**, 851–855 (2002).
57. Kelly, J. W. & Balch, W. E. Amyloid as a natural product. *J. Cell Biol.* **161**, 461–462 (2003).
58. Broome, B. M. & Hecht, M. H. Nature disfavors sequences of alternating polar and non-polar amino acids: implications for amyloidogenesis. *J. Mol. Biol.* **296**, 961–968 (2000).
59. Chiti, F. *et al.* Kinetic partitioning of protein folding and aggregation. *Nature Struct. Biol.* **9**, 137–143 (2002).
60. Macario, A. J. L. & Macario, E. C. Sick chaperones and ageing: a perspective. *Ageing Res. Rev.* **1**, 295–311 (2002).
61. Ramirez-Alvarado, M., Merkel, J. S. & Regan, L. A systematic exploration of the influence of the protein stability on amyloid fibril formation in vitro. *Proc. Natl Acad. Sci. USA* **97**, 8979–8984 (2000).
62. Dumoulin, M. *et al.* A camelid antibody fragment inhibits amyloid fibril formation by human lysozyme. *Nature* **424**, 783–788 (2003).
63. Prusiner, S. B. Prion diseases and the BSE crisis. *Science* **278**, 245–251 (1997).
64. Taylor, J. P., Hardy, J. & Fischbeck, K. H. Toxic proteins in neurodegenerative disease. *Science* **296**, 1991–1995 (2002).
65. Walsh, D. M. *et al.* Naturally secreted oligomers of amyloid β protein potently inhibit hippocampal long-term potentiation *in vivo*. *Nature* **416**, 535–539 (2002).
66. Kaye, R. *et al.* Common structure of soluble amyloid oligomers implies common mechanisms of pathogenesis. *Science* **300**, 486–489 (2003).
67. Stefani, M. & Dobson, C. M. Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution. *J. Mol. Med.* **81**, 678–699 (2003).
68. Sherman, M. Y. & Goldberg, A. L. Cellular defenses against unfolded proteins: a cell biologist thinks about neurodegenerative diseases. *Neuron* **29**, 15–32 (2001).
69. Muchowski, P. J. *et al.* Hsp70 and Hsp40 chaperones can inhibit self-assembly of polyglutamine proteins into amyloid-like fibrils. *Proc. Natl Acad. Sci. USA* **97**, 7841–7846 (2000).
70. Csermely, P. Chaperone overload is a possible contributor to 'civilization diseases'. *Trends Genet.* **17**, 701–704 (2001).
71. Dobson, C. M. Getting out of shape—protein misfolding diseases. *Nature* **418**, 729–730 (2002).

Acknowledgements I should like to thank in particular the Wellcome Trust, the Leverhulme Trust and the UK Research Councils for generous support over many years, without whom my own research activities in this area of science could not have been carried out.

Quality control in the endoplasmic reticulum protein factory

Roberto Sitia¹ & Ineke Braakman²

¹Università Vita-Salute San Raffaele Scientific Institute, 20132 Milano, Italy (e-mail: r.sitia@hsr.it)

²Bio-Organic Chemistry 1, Utrecht University, 3584 CH Utrecht, The Netherlands (e-mail: i.braakman@chem.uu.nl)

The endoplasmic reticulum (ER) is a factory where secretory proteins are manufactured, and where stringent quality-control systems ensure that only correctly folded proteins are sent to their final destinations. The changing needs of the ER factory are monitored by integrated signalling pathways that constantly adjust the levels of folding assistants. ER chaperones and signalling molecules are emerging as drug targets in amyloidoses and other protein-conformational diseases.

A aberrant proteins are extremely harmful to cells. Such proteins can originate from mutations, unbalanced subunit synthesis, damaging conditions (for example, oxidation) and also as a side product of normal protein biosynthesis. A certain degree of folding inefficiency can be useful. Misfolded proteins are degraded into fragments (peptides), which are then displayed on the cell's surface. In cells that are not infected, normal cytosolic and nuclear proteins will be displayed; in cells that are infected with a virus, viral peptides are presented and then detected by immune cells. The presence of membrane-bounded compartments in eukaryotic cells further complicates the issue of protein homeostasis, as it implies the existence and coordination of multiple folding and degradation systems.

A proper balance between synthesis, maturation and degradation is crucial for cell survival. At a time when health threats come from viruses, misfolded proteins (which can cause Creutzfeldt–Jakob disease) and inherited diseases, and when protein-based drugs are being introduced, a full understanding of how cells handle proteins is of crucial importance.

In eukaryotic cells, the vast majority of proteins fold either in the cytosol or in the ER. Although the general principles are similar in the both compartments^{1,2}, fundamental differences exist. Here we summarize the main features of folding and quality control in the ER, highlighting emerging concepts of how the functions of the ER are integrated with other cellular compartments and with transcriptional and translational control to prevent proteotoxicity.

Protein folding in the ER

The ER lumen is similar to the extracellular space: it has high calcium concentrations and is more oxidizing than the cytosol (see refs 1, 2 and references therein), and contains a specialized set of chaperones and enzymes (Table 1). The structural maturation of many proteins synthesized in the ER is slow and inefficient¹, probably because they require several post-translational modifications (for example, signal sequence cleavage, N-linked glycosylation, disulphide-bond formation and reshuffling, addition of glycosylphosphatidylinositol anchors, insertion of membrane proteins in the lipid bilayer). Coordinating these covalent modifications is a challenging task for the folding machinery in the ER. Nevertheless, the ER factory can reach levels of extraordinary efficiency in exocrine glands, plasma cells and other 'professional secretory cells'.

In mammalian cells, proteins are translocated into the

ER (Fig. 1), where they start to fold co-translationally. Folding is completed post-translationally, and, generally, individual subunits have folded before assembly and oligomerization take place¹. Sequential interactions with distinct chaperones are required for each of these steps. For instance, MHC class I molecules exploit calnexin, calreticulin and other devoted ER-resident molecules (for example, tapasin) to negotiate their stepwise assembly into peptide-loaded functional complexes³. Although most folding factors in the ER are ubiquitously expressed throughout the body, some are tissue-type specific or cell-type specific, and probably fulfil a particular synthetic task (Table 1). For example, efficient collagen production requires the expression of hsp47, whereas a tissue-specific protein-disulphide-isomerase-like protein, PDIP, is produced in the pancreas and probably permits the massive secretion of digestive enzymes^{1,4}.

Disulphide-bond formation

Both N-glycosylation and disulphide-bond formation play a crucial role in the folding of secretory and membrane proteins in the ER. The glycans add hydrophobicity to the folding peptide but also act as substrates for the lectin chaperones calnexin and calreticulin, whose role in glycoprotein folding has been extensively reviewed^{5,6}. Both calnexin and calreticulin form a complex with ERp57, an ER oxidoreductase, coupling folding assistance to disulphide-bond formation.

The impressive number of oxidoreductases in the ER suggests that catalysis and regulation of disulphide-bond formation is crucial for folding. Energywise, in most cases, the contribution of a disulphide bond is hardly more than that of a single hydrogen bond, yet, without disulphide bonds, native conformations are not obtained. Disulphide bonds cannot force a folding protein into a given conformation: in the sampling of conformations during folding in the ER, native and non-native disulphide cross-links are transiently formed. Continuous activity of oxidoreductases probably ensures that these covalent links remain flexible until folding is completed.

Secreted proteins enter the ER with reduced cysteines and leave it with oxidized cystines. The requirement for oxidative equivalents is fulfilled by Ero1 proteins, which transfer electrons from protein disulphide isomerase (PDI) to oxygen through a series of specific interchange reactions^{7,8}. Additional electron transport pathways exist, and involve proteins such as Erv2p from yeast, although their role under normal conditions remains to be deter-

mined⁷. The redox gradient between the ER and the cytosol seems to be important for intercompartmental signalling, particularly in the integrated response to oxidative stress, in which adaptive responses emanating from different compartments are coordinated⁹. And redox reactions with opposite electron fluxes must take place in the ER to mediate formation, isomerization and reduction of disulphides⁹. The wealth of redox assistants allows these fluxes to be separate, and channels electron transport through specific protein–protein interactions. The main role of glutathione in redox homeostasis in the ER seems to be that of buffering the oxidative power of Ero1 (refs 7, 8). Because disulphide bonds are so important for folding, we may conclude from the number of ER-resident oxidoreductases that secretory proteins need more help, possibly because they are often larger than cytosolic ones¹⁰. Perhaps secretory proteins, which are designed to act extracellularly and often must carry their biological messages over long distances, need more protection from denaturing forces outside the cell. This is an issue that glycans are likely to contribute to as well.

Quality control in the ER

Besides providing a unique folding environment, the ER has a crucial quality-control role^{1,2,4}. How does it discriminate between native and non-native proteins? The answer to this question depends primarily on ER chaperones. When folding or assembly intermediates expose hydrophobic surfaces, unpaired cysteines or immature glycans, ER-resident chaperones or oxidoreductases interact with them, and as a

consequence they are retained in the ER or retrieved from the Golgi complex¹⁴. By forming multimolecular complexes¹¹, folding factors in the ER may provide matrices that couple retention to folding and assembly. Immature proteins may also form aggregates that are excluded from vesicles exiting from the ER^{1,12}.

All proteins are subjected to a 'primary' quality control that monitors their architectural design through ubiquitous folding sensors (Table 1). 'Secondary' quality-control mechanisms rely instead on cell-specific factors and facilitate export of individual proteins or classes of proteins (for example, tapasin for class I molecules; see also refs 1, 2, 4 and 13 and references therein for additional examples). Although some proteins can be rerouted to and degraded in the endolysosomal compartment (Fig. 2), the ER is the main test bench where molecules destined for the extracellular space are scrutinized for their potential toxicity. Interestingly, ER quality control seems to monitor primarily local structures within protein domains, allowing transport of proteins originating from exon reshuffling or otherwise 'flexible' molecules¹⁴. A certain degree of freedom from quality control is essential for the evolution of proteins. However, it comes at a price for multicellular organisms. Indeed, many proteins that cause systemic amyloidosis (see the review in this issue by Dobson, page 884) can adopt more than one conformation and can undergo uncontrolled aggregation outside of the cells.

The reasons for having a quality-control system in the ER are easy to understand where protein folding and function are concerned, especially in multicellular organisms where development relies on the fidelity of protein secretion. Quality control can also regulate the transport or the activity of certain proteins during differentiation¹⁴ or in response to stress or metabolic requirements. A clear case in the context of ER physiology is that of the transcription factor ATF6, a transmembrane protein localized in the ER by interactions between its luminal domain and BiP, an abundant ER-resident chaperone of the hsp70 family (Table 1). The unfolded proteins titrate BiP, releasing ATF6 for transport and subsequent cleavage by Golgi proteases^{15–17}. In this way, the active cytosolic domain is unleashed and delivered to the nucleus, where it drives transcription of target genes. A similar cleavage mechanism to that for ATF6 underlies the cholesterol-dependent transport of sterol regulatory element binding protein¹⁸. Other cases of regulated transport are based on the availability of specific ligands (for example, lipids and retinol) or cofactors (calcium ions, vitamin C and so on). Clarifying the mechanisms controlling differential ER to Golgi transport has obvious technological and clinical implications. For example, ligand-induced transport may allow storage of hormones or other proteins in the ER, which can be released when necessary by ligand administration¹⁹.

ER-associated degradation

Mutations or unbalanced subunit synthesis make folding or assembly — and hence exit from the ER — impossible. To maintain homeostasis, terminally misfolded molecules are 'retrotranslocated' or 'dislocated' across the ER membrane to be degraded by cytosolic proteasomes^{20,21}. Dislocation is generally believed to occur through Sec61, a protein complex also used by nascent proteins to enter into the ER (Fig. 1), raising the question of how vectorial transport in opposite directions is regulated. Additional mechanisms for delivering molecules across the ER membrane may also exist.

A fascinating problem is how molecules that have not been given the time to fold (and therefore are unfolded) are discriminated from those that have failed to fold after many attempts (misfolded), and must therefore be disposed of. One way of timing glycoprotein quality control involves the sequential processing of N-glycans and in particular mannose trimming in the ER²². It remains to be seen how substrates are eventually targeted to the retrotranslocation channels, how these are opened, and to what extent proteins must be unfolded to negotiate dislocation^{8,21}.

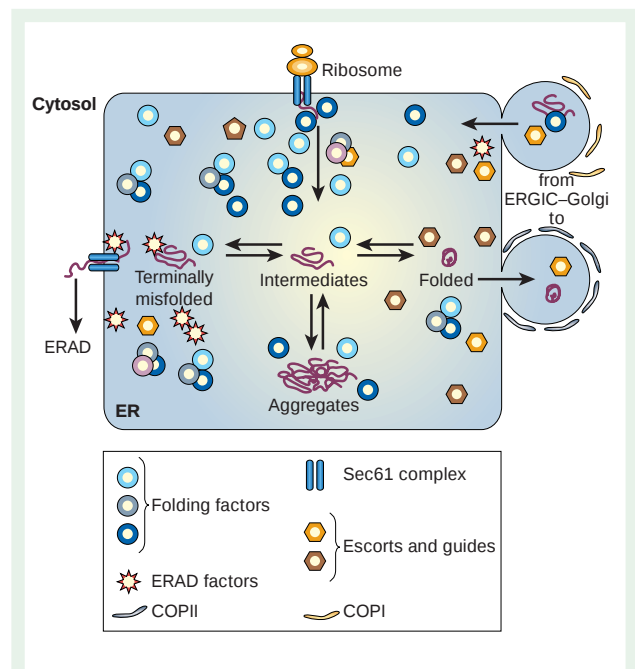


Figure 1 Folding in the crowded environment of the ER. The port of entry for proteins destined for the secretory pathway is the ER. These proteins are synthesized by ER-associated ribosomes and co-translationally translocated across the membrane through the Sec61 complex. The ER is rich in chaperones and folding enzymes (folding factors), in molecules involved in mediating transport to the cytosol for proteasomal degradation (ERAD factors) or to the downstream stations of the secretory pathway (escort and guides¹³). Some ER-resident proteins seem to form multimolecular complexes¹, which can be excluded from transport by size selection, and provide a matrix that couples retention to folding¹⁴. A distinctive feature of the ER is its ability to catalyse opposite reactions: folding and unfolding, oxidation and reduction, protein import and export through Sec61. It is debated whether a 'quality-control compartment'^{12,39}, involved in the recognition and targeting of terminally misfolded proteins (alluded to in the picture by having ERAD factors concentrated on the left), exists. ERGIC, ER–Golgi intermediate compartment.

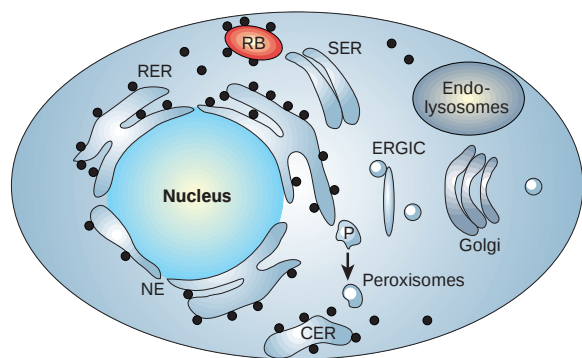


Figure 2 The ER is not always that rough. Distinct ER subcompartments can be clearly identified: the nuclear envelope (NE), and rough and smooth stacks (RER and SER). In plant cells, dilated ER cisternae are used for protein storage. In mammalian cells, they are generally associated with the accumulation of mutated or transport-incompetent proteins produced in excess with respect to ERAD capacity (for example, Russell bodies, RB³⁰). Additional ER subdomains were identified recently, including a peroxisome precursor compartment (P) and the cortical ER (CER) in yeast^{40,41}. Selected mRNAs seem to be targeted to the storage domains of plant ERs as well as to the yeast CER, thus favouring subcompartmentalization. ER subdomains could exist at an even smaller scale¹¹. For instance, different complexes of ER folding factors in the ER may fulfil sequential (and often opposite) tasks during protein maturation in the ER, and membrane sub-domains might exist.

The unfolded protein response and signalling from the ER

To maintain the efficiency of quality-control mechanisms in diverse physiological conditions, living cells have evolved regulatory circuits that monitor the levels of available chaperones. This is true for both the cytosol and the ER, and compartment-specific responses clearly exist that selectively restore optimal levels of the desired folding factors. The accumulation of aberrant proteins in the cytosol triggers the heat-shock response, resulting in *de novo* synthesis of hsp70 and other cytosolic chaperones²³. But if aberrant proteins accumulate in the ER, cells activate a different response, the unfolded protein response (UPR), which leads to the coordinated synthesis of ER-resident chaperones and enzymes. The UPR is induced by drugs that block N-linked glycosylation or disulphide-bond formation, or alter calcium ion homeostasis, thus selectively targeting the ER folding machinery^{16,17}. Physiologically, ER stress (a condition in which the folding machinery in the ER cannot cope with its protein load) can be caused by synthesis of mutated or orphan proteins, absence of cofactors (an example being scurvy, in which collagen cannot fold because of the lack of vitamin C²⁴), or a drastic increase in otherwise normal cargo proteins.

How do the diverse unfolded or misfolded proteins that accumulate in the ER provoke the same pathway? The unifying concept is that BiP and other primary quality-control factors maintain the stress sensors in the ER in the inactive state^{16,17}, so that chaperone insufficiency triggers UPR whatever the nature of the cargo.

The mammalian ER sensors, Ire1, PERK and ATF6, guarantee a tripartite response with synergic strategies^{16,17}. By phosphorylating eIF2 α , PERK transiently attenuates translation, limiting protein load. ATF6 drives the transcriptional upregulation of many ER-resident proteins and folding assistants. Ire1 activates XBP-1, which in turn induces transcription of factors that facilitate ER-associated degradation (ERAD). The two-step activation of XBP-1 (transcriptionally induced by ATF6 and post-transcriptionally regulated by Ire1) guarantees the proper timing of the UPR; attempts to fold proteins precede the decision to degrade them²⁵. If the response fails to

Table 1 Personnel of the ER protein factory in mammalian cells

Ubiquitous ER-resident proteins

Family	Main members	Functions
Hsp70	BiP/grp78	Chaperone
Hsp90	Endoplasmic/grp94	Chaperone
Hsp40	ERdj1-ERdj5, Sec63	Co-chaperones regulating BiP?
Lectins	Calnexin and calreticulin	Glycoprotein quality control
	EDEM	Glycoprotein degradation ²²
Glycan-processing enzymes	UGT	Folding sensor. Adds glucoses to misfolded glycoproteins (enters substrates into the calnexin cycle) ^{5,6}
	ER glucosidases I and II	Remove glucoses from N-glycans (on/off/calnexin cycle) ^{5,6}
	ER mannosidases I and II	Remove terminal mannoses — ERAD? ²²
Peptidyl-prolyl isomerases	Cyclophilins, FK506-binding proteins	Isomerize <i>cis-trans</i> peptidyl-prolyl bonds?
Ero1	Ero1 α , Ero1 β	Disulphide-bond formation ^{7,8}
Oxidoreductases	PDI, ERp72, ERp57, p5, and many others	Disulphide-bond formation, isomerization and reduction

Examples of specific folding assistants

Name	Tissue distribution	Function
PDIp	Pancreas	Oxidoreductase
Hsp47 and prolyl-4-hydroxylase	Cells producing collagen	Collagen synthesis and assembly ²⁴
Invariant chain	Antigen presenting cells	MHC class II assembly and transport
Tapasin	Ubiquitous	Peptide binding to MHC class I ³
RAP	Ubiquitous	LDL receptor assembly and transport
BOCA/Mesd	?	LDL receptor assembly and transport ⁴²
Egagyn	Ubiquitous	Quality control of β -glucuronidase
Carboxylesterase	Hepatocytes	Quality control of C reactive proteins
SCAP	Ubiquitous	Retention of SREBP ¹⁸

clear the ER, apoptosis is induced through several pathways^{16,17,26}. This link has important consequences for the pathogenesis of degenerative disorders (see the progress in this issue by Goldberg, page 895).

The UPR is multi-faceted and regulates proteins involved in quality control, ERAD and many aspects of the secretory pathway²⁷. It is emerging as a key controller of normal development as well. Particularly compelling in this respect is the observation that plasma-cell differentiation requires XBP-1 (ref. 28). However, attenuating translation through PERK would be counterproductive for plasma cells, which must release antibodies in large quantities²⁶. It remains to be seen whether cells can selectively activate individual branches of the UPR. One possibility is that the different ER-resident proteins containing a DnaJ motif (ERdj)²⁹ might preferentially release BiP from one of the three UPR sensors, similar to the way in which E3 ligases confer specificity on E1 and E2 proteins.

ER quality control and disease

Quality control must be a balance between retaining and degrading potentially harmful products and not preventing export of biologically active proteins. CFTR mutants in cystic fibrosis illustrate an overzealous quality control, where biologically active mutants cannot leave the ER. In this case, relaxing the quality control could cure the patient. But disease can also originate from defective degradation. If the rate of synthesis of a protein exceeds the combined rates of folding and degradation, a fraction of it will accumulate intracellularly. At least two bottlenecks are encountered by ERAD substrates: dislocation across the membrane and actual degradation by the proteasome. Inefficient proteolysis often results in the formation of deposits of ubiquitylated proteins in the pericentriolar region, called aggresomes³⁰. Studies with the prion protein suggest that mislocalization to the cytosol is sufficient to cause conversion into the PrP^{Sc} form^{31,32}. By contrast, inefficient dislocation results in substrates accumulating in the ER. Many ER storage diseases are characterized by the presence of dilated ER cisternae containing mutated pro-

teins³⁰. It is not clear whether these aberrant structures are toxic *per se* or whether they represent a defence mechanism that segregates dangerous proteins into specialized ER subcompartments (Fig. 2). The tendency of the mutated protein to sequester chaperones, and hence to induce a prolonged UPR leading to apoptosis, is probably important in determining cell damage.

Proteasome inhibitors are powerful inducers of the UPR, probably because dislocation of many substrates is coupled to degradation. Receiving substrates from both ER and cytosol, proteasomes are crossroads of the two main protein quality-control systems. Conditions that hamper proteasome function, such as the presence of proteins with polyglutamine repeats^{33,34} or a lack of E3 ligases³⁵, therefore can cause accumulation of ERAD substrates in the ER and activate UPR-dependent apoptotic pathways. In this way, proteotoxicity could be transmitted across compartments, with important implications for degenerative diseases (reviewed in refs 35, 36).

Perspectives

Over the past few years, much has been learned about how proteins are handled by the ER folding and quality-control machineries, and some of this knowledge has begun to be translated to industry and to the clinic. Yet, many questions remain: how does aggregation relate to degradation? Are aggregation and degradation sometimes mutually exclusive, with aggregation preventing dislocation? In what circumstances does this happen? How does aggregation relate to amyloid formation, and how can aggregates be eliminated? How do cells secrete large molecular ensembles, such as procollagen and viruses? Even before complete answers are provided, new strategies for intervention are being considered to prevent viral replication, protein aggregation or amyloid formation. The most important question is whether the folding of a protein can be influenced specifically, with the goals either of curing a disease or of preventing virus assembly. Chemical chaperones and ligand-induced transport opens options for designing specific drugs to control protein (mis)folding or transport^{37,38}. Likewise, tissue-specific chaperones might be therapeutic targets and might provide important tools in biotechnology. Will we be able to reduce protein misfolding in degenerative disorders, or induce it in some tumours so as to cause apoptosis? Last but not least, the UPR signalling cascades clearly offer targets for pharmacological intervention. Drugs that modulate the UPR may alleviate a variety of age-related diseases resulting from protein misfolding and aggregation. □

doi:10.1038/nature02262

1. Ellgaard, L. & Heliénus, A. Quality control in the endoplasmic reticulum. *Nature Rev. Mol. Cell Biol.* **4**, 181–191 (2003).
2. Fewell, S. W., Travers, K. J., Weissman, J. S. & Brodsky, J. L. The action of molecular chaperones in the early secretory pathway. *Annu. Rev. Genet.* **35**, 149–191 (2001).
3. Bouvier, M. Accessory proteins and the assembly of human class I MHC molecules: a molecular and structural perspective. *Mol. Immunol.* **39**, 697–706 (2003).
4. Ellgaard, L., Molinari, M. & Heliénus, A. Setting the standards: quality control in the secretory pathway. *Science* **286**, 1882–1888 (1999).
5. Schrag, J. D., Procopio, D. O., Cygler, M., Thomas, D. Y. & Bergeron, J. J. Lectin control of protein folding and sorting in the secretory pathway. *Trends Biochem. Sci.* **28**, 49–57 (2003).
6. Heliénus, A. & Aebi, M. Intracellular functions of N-linked glycans. *Science* **291**, 2364–2369 (2001).
7. Sevier, C. S. & Kaiser, C. A. Formation and transfer of disulphide bonds in living cells. *Nature Rev. Mol. Cell Biol.* **3**, 836–847 (2002).
8. Fassio, A. & Sitia, R. Formation, isomerisation and reduction of disulphide bonds during protein quality control in the endoplasmic reticulum. *Histochem. Cell Biol.* **117**, 151–157 (2002).
9. Harding, H. P. *et al.* An integrated stress response regulates amino acid metabolism and resistance to oxidative stress. *Mol. Cell* **11**, 619–633 (2003).
10. van Anken, E. *et al.* Sequential waves of functionally related proteins are expressed when B cells prepare for antibody secretion. *Immunity* **18**, 243–253 (2003).

11. Meunier, L., Usherwood, Y. K., Chung, K. T. & Hendershot, L. M. A subset of chaperones and folding enzymes form multiprotein complexes in endoplasmic reticulum to bind nascent proteins. *Mol. Biol. Cell* **13**, 4456–4469 (2002).
12. Hendershot, L. M. Giving protein traffic the green light. *Nature Cell Biol.* **2**, E105–E106 (2000).
13. Herrmann, J. M., Malkus, P. & Schekman, R. Out of the ER — outfitters, escorts and guides. *Trends Cell Biol.* **9**, 5–7 (1999).
14. Sitia, R. *et al.* Developmental regulation of IgM secretion: the role of the carboxy-terminal cysteine. *Cell* **60**, 781–790 (1990).
15. Shen, J., Chen, X., Hendershot, L. & Prywes, R. ER stress regulation of ATF6 localization by dissociation of BiP/GRP78 binding and unmasking of Golgi localization signals. *Dev. Cell* **3**, 99–111 (2002).
16. Kaufman, R. J. Orchestrating the unfolded protein response in health and disease. *J. Clin. Invest.* **110**, 1389–1398 (2002).
17. Harding, H. P., Calton, M., Urano, F., Novoa, I. & Ron, D. Transcriptional and translational control in the mammalian unfolded protein response. *Annu. Rev. Cell Dev. Biol.* **18**, 575–599 (2002).
18. Brown, A., Sun, L., Feramisco, J., Brown, M. & Goldstein, J. Cholesterol addition to ER membranes alters conformation of SCAP, the SREBP escort protein that regulates cholesterol metabolism. *Mol. Cell* **10**, 237–245 (2002).
19. Rivera, V. M. *et al.* Regulation of protein secretion through controlled aggregation in the endoplasmic reticulum. *Science* **287**, 826–830 (2000).
20. Tsai, B., Ye, Y. & Rapoport, T. A. Retro-translocation of proteins from the endoplasmic reticulum into the cytosol. *Nature Rev. Mol. Cell Biol.* **3**, 246–255 (2002).
21. Kostova, Z. & Wolf, D. H. For whom the bell tolls: protein quality control of the endoplasmic reticulum and the ubiquitin-proteasome connection. *EMBO J.* **22**, 2309–2317 (2003).
22. Wang, T. & Hebert, D. N. EDEM an ER quality control receptor. *Nature Struct. Biol.* **10**, 319–321 (2003).
23. Morimoto, R. I. Regulation of the heat shock transcriptional response: cross talk between a family of heat shock factors, molecular chaperones, and negative regulators. *Genes Dev.* **12**, 3788–3796 (1998).
24. Hosokawa, N. & Nagata, K. Procollagen binds to both prolyl 4-hydroxylase/protein disulfide isomerase and HSP47 within the endoplasmic reticulum in the absence of ascorbate. *FEBS Lett.* **466**, 19–25 (2000).
25. Yoshida, H. *et al.* A time-dependent phase shift in the mammalian unfolded protein response. *Dev. Cell* **4**, 265–271 (2003).
26. Ma, Y. & Hendershot, L. M. The unfolding tale of the unfolded protein response. *Cell* **107**, 827–830 (2001).
27. Higashio, H. & Kohno, K. A genetic link between the unfolded protein response and vesicle formation from the endoplasmic reticulum. *Biochem. Biophys. Res. Commun.* **296**, 568–574 (2002).
28. Reimold, A. M. *et al.* Plasma cell differentiation requires the transcription factor XBP-1. *Nature* **412**, 300–307 (2001).
29. Cunniff, P. M. *et al.* ERdj5, an endoplasmic reticulum (ER)-resident protein containing DnaJ and thioredoxin domains, is expressed in secretory cells or following ER stress. *J. Biol. Chem.* **278**, 1059–1066 (2003).
30. Kopito, R. R. & Sitia, R. Aggregates and Russell bodies. *EMBO Rep.* **1**, 225–231 (2000).
31. Ma, J., Wollmann, R. & Lindquist, S. Neurotoxicity and neurodegeneration when PrP accumulates in the cytosol. *Science* **298**, 1781–1785 (2002).
32. Hegde, R. S. & Rane, N. S. Prion protein trafficking and the development of neurodegeneration. *Trends Neurosci.* **26**, 337–339 (2003).
33. Nishitoh, H. *et al.* ASK1 is essential for endoplasmic reticulum stress-induced neuronal cell death triggered by expanded polyglutamine repeats. *Genes Dev.* **16**, 1345–1355 (2002).
34. Bence, N. F., Sampat, R. M. & Kopito, R. R. Impairment of the ubiquitin-proteasome system by protein aggregation. *Science* **292**, 1552–1555 (2001).
35. Imai, Y. *et al.* An unfolded putative transmembrane polypeptide, which can lead to endoplasmic reticulum stress, is a substrate of Parkin. *Cell* **105**, 891–902 (2001).
36. Yoneda, T., Urano, F. & Ron, D. Transmission of proteotoxicity across cellular compartments. *Genes Dev.* **16**, 1307–1313 (2002).
37. Brown, C. R., Hong-Brown, L. Q. & Welch, W. J. Correcting temperature-sensitive protein folding defects. *J. Clin. Invest.* **99**, 1432–1444 (1997).
38. Aridor, M. & Balch, W. E. Perspectives: drug delivery. Regulating export of ER cargo. *Science* **287**, 816–817 (2000).
39. Kamhi-Nesher, S. *et al.* A novel quality control compartment derived from the endoplasmic reticulum. *Mol. Biol. Cell* **12**, 1711–1723 (2001).
40. Voeltz, G. K., Rolls, M. M. & Rapoport, T. A. Structural organization of the endoplasmic reticulum. *EMBO Rep.* **3**, 944–950 (2002).
41. Papp, S., Dziak, E., Michalak, M. & Opat, M. Is all of the endoplasmic reticulum created equal? The effects of the heterogeneous distribution of endoplasmic reticulum Ca²⁺-handling proteins. *J. Cell Biol.* **160**, 475–479 (2003).
42. Herz, J. & Marschang, P. Coaxing the LDL receptor family into the fold. *Cell* **112**, 289–292 (2003).

Acknowledgments We thank members of our laboratories for their criticisms and enthusiasm, D. Ron for helpful suggestions, T. Mastrandrea for secretarial assistance, and Associazione Italiana per la Ricerca sul Cancro (AIRC), Italian Ministries of Health, Education and Research (CoFin and Center of Excellence in Physiopathology of Cell Differentiation), NWO and Telethon for financial support. We apologize to the many colleagues whose seminal work we could not cite because of space limitations.

Protein degradation and protection against misfolded or damaged proteins

Alfred L. Goldberg

Department of Cell Biology, Harvard Medical School, Boston, Massachusetts 02115, USA
(e-mail: alfred_goldberg@hms.harvard.edu)

The ultimate mechanism that cells use to ensure the quality of intracellular proteins is the selective destruction of misfolded or damaged polypeptides. In eukaryotic cells, the large ATP-dependent proteolytic machine, the 26S proteasome, prevents the accumulation of non-functional, potentially toxic proteins. This process is of particular importance in protecting cells against harsh conditions (for example, heat shock or oxidative stress) and in a variety of diseases (for example, cystic fibrosis and the major neurodegenerative diseases). A full understanding of the pathogenesis of the protein-folding diseases will require greater knowledge of how misfolded proteins are recognized and selectively degraded.

Proteins within cells are continually being degraded to amino acids and replaced by newly synthesized proteins. This process is highly selective and precisely regulated¹, and individual proteins are destroyed at widely different rates, with half-lives ranging from several minutes to many days. In eukaryotic cells, most proteins destined for degradation are labelled first by ubiquitin in an energy-requiring process and then digested to small peptides by the large proteolytic complex, the 26S proteasome. Indicative of the complexity and importance of this system is the large number of gene products (perhaps a thousand) that function in the degradation of different proteins in mammalian cells. In the past decade, there has been an explosion of interest in the ubiquitin–proteasome pathway, due largely to the general recognition of its importance in the regulation of cell division, gene expression and other key processes¹. However, the cell's degradative machinery must have evolved initially to serve a more fundamental homeostatic function — to serve as a quality-control system that rapidly eliminates misfolded or damaged proteins whose accumulation would interfere with normal cell function and viability^{2–4}.

Long before the appearance of the ubiquitin–proteasome pathway in eukaryotes, prokaryotes had evolved elaborate proteolytic machinery to destroy misfolded proteins rapidly^{4,5}. These intracellular proteolytic systems are energy dependent and differ markedly in size and complexity from the simple proteases that function in the extracellular milieu, such as the pancreatic proteases that digest dietary proteins. If such typical extracellular proteases were free in the cytosol, they would quickly convert the cell into a bag of amino acids. The fundamental problem that evolution had to solve was how to provide cells with the capacity to destroy misfolded or damaged proteins rapidly without the non-specific destruction of essential cell constituents.

Here, I outline the key features of the degradative process, its importance for cell viability and the research that led to its elucidation. As discussed here and elsewhere in this Insight, misfolded proteins can arise in cells by mutation and through various postsynthetic events; they can be highly toxic, mainly because of their tendency to form intracellular aggregates. Greater knowledge about this proteolytic pathway will be crucial for understanding the pathogenesis of diseases resulting from failures in correct protein folding and in developing rational therapies.

Structural alterations lead to rapid degradation

The ability of cells to degrade abnormally folded proteins selectively was first demonstrated more than 30 years ago⁴ with the discovery that various treatments that perturb the proper folding of proteins caused their rapid hydrolysis. In fact, these findings led to the initial recognition that a protein's structure determines not only its catalytic properties but also its intracellular stability². However, the precise changes in conformation that are recognized by the cell's degradative machinery and that trigger rapid hydrolysis are not clear and have still not been studied systematically.

Perhaps most illustrative are the early findings about degradation of abnormal globins in reticulocytes^{6,7}, which produce one very well-characterized protein, haemoglobin, almost exclusively. Because the powerful tools that are now available to perturb protein structures (such as site-directed mutagenesis) were undreamt of then, those early studies followed the fate of proteins that had incorporated, in place of the natural residues, amino-acid analogues that prevented the protein from assuming its normal tertiary conformation^{2,4}. Normally, haemoglobin is the most stable intracellular protein, lasting the lifespan of red cells (about 110 days). However, after the incorporation of a synthetic valine analogue, the abnormal globin had a half-life of 10 min (refs 6, 7). These short-lived globins failed to bind haem or form tetramers, and instead formed amorphous aggregates before degradation (see below). Very similar events have since been observed in several human diseases, for example, the unstable haemoglobinopathies, in which mutations that prevent the binding of haem cause complete degradation within minutes of translation, leading to a severe anaemia⁸.

Wider recognition of the importance of this degradative process occurred with the advent of recombinant DNA technology in the 1980s and the discovery that many foreign proteins expressed in bacteria fail to accumulate owing to their rapid degradation. Many proteins of major medical importance (such as insulin) cannot fold properly in *Escherichia coli*, are recognized as abnormal and are rapidly degraded⁹. Successful expression of such proteins required the development of mutant bacteria with reduced degradative capacity or improved expression systems that overwhelm the proteolytic systems⁹.

Although most of our knowledge about misfolded proteins comes from mutated proteins, cells also rapidly degrade abnormal proteins that result from errors in transcription or translation. Accordingly, ribosomal mutations

Table 1 Abnormal proteins rapidly degraded in cells

Type of abnormality	Cause
Incomplete proteins	Nonsense mutations, incorporation of puromycin, premature termination, proteolytic cleavage
Missense proteins	Mutations, incorporation of amino-acid analogues, biosynthetic errors
Free subunits of multimeric complexes	Excess synthesized subunits
Postsynthetic damage	Oxygen radicals, intracellular denaturation
Genetic engineering	Gene fusions, frame-shifts, incorrect localization
Protein misfolding	

in *E. coli* or treatment with streptomycin, each of which causes translational errors and the production of error-laden proteins, leads to an overall increase in proteolysis^{4,11}. In fact, because of this rapid degradation, the actual error rate in gene expression *in vivo* is impossible to determine with precision.

Incomplete proteins, such as can arise through mistakes in RNA splicing, the incorporation of puromycin^{2,4,6} or nonsense mutations, are rapidly degraded in all cells. Such fragments fail to assume their proper conformation. Bacteria also have a specialized mechanism to target for destruction incomplete polypeptides while they are still on stalled ribosomes¹¹, but no similar mechanism has yet been found in eukaryotes.

The quality-control systems in bacteria and eukaryotic cells do not degrade all, or even most, mutated proteins: only those mutations that markedly perturb protein folding trigger rapid hydrolysis (such as mutations of key residues or large deletions) (see Table 1). For example, only about one-fifth of the hundreds of human haemoglobin point mutations generate proteins that undergo rapid degradation (as suggested by the low levels of the abnormal protein in mature red cells)⁸. A major recent development has been the discovery that many abnormally folded proteins in the secretory pathway (for example, the cystic fibrosis transmembrane conductance regulator) are translocated back into the cytosol, where they undergo rapid degradation (see the article by Sitia and Braakman in this issue p. 891, and ref. 10).

Degradation of newly synthesized proteins

The folding of newly synthesized proteins to their proper conformations involves the sequential actions of multiple molecular chaperones^{13,14}, and this process can take many minutes or even longer, and may often be unsuccessful. A large fraction of newly synthesized proteins is rapidly degraded^{2,15,16}. As many as 30% of newly synthesized proteins in eukaryotes might undergo degradation within minutes of synthesis¹⁵. Although initially attributed to mistakes in ribosomal function (and named DRIPS, for 'defective ribosomal products'), it seems more likely that a large fraction of these short-lived species are products of unsuccessful folding or failures of multimer assembly. However, the actual fraction of short-lived species that are genuinely aberrant proteins is controversial and extremely difficult to determine rigorously for multiple reasons. For example, a number of important regulatory proteins have very short half-lives (as short as 2–20 min)¹; also fragments of all secretory and membrane proteins (signal peptides) are degraded as part of the secretory process. Nevertheless, it seems likely that many newly synthesized polypeptides are destroyed because of the inherent inefficiency of protein folding¹⁵ and that, after release from the ribosome, polypeptides face a life–death kinetic competition between successful folding and rapid hydrolysis of unfolded species.

Particularly critical steps in the folding process are the binding of cofactors and the association of different subunits to form multimeric complexes^{13,14}. In the absence of cofactors or complementary subunits, free subunits generally fail to achieve native conformations and have exposed hydrophobic surfaces. Such free subunits tend to be

rapidly degraded². For example, ribosomal subunits, when not in ribosomes, are very unstable in all cells. In human thalassaemia, there is an excess of α - and β -globin subunits not in tetramers that are rapidly degraded within reticulocytes⁸. However, their destruction is not sufficiently rapid to prevent free chains from precipitating and distorting erythrocyte shape. This defect in protein assembly leads eventually to anaemia because these misshapen red cells are rapidly destroyed by the spleen (which functions as a quality-control system to ensure that circulating cells have the correct shape)⁸. The degradation of free subunits of multimeric enzymes thus not only protects cells against accumulation of denatured, potentially toxic polypeptides but also functions as a quality-control mechanism that ensures that subunits of multimeric complexes are present in the proper stoichiometry.

Postsynthetic damage to cell proteins

For a protein, the cell can hardly be a hospitable environment, especially in warm-blooded species. Within cells, proteins are constantly exposed to highly reactive molecules and to conditions that favour denaturation (see Box 1) and are under constant surveillance by the proteolytic systems, which continually monitor mature proteins for postsynthetic denaturation or chemical damage. Consequently, a protein's life within the cell is likely to be 'nasty, brutish and short'.

To maintain the activity of most proteins in the laboratory, it is essential to store them at low temperatures in pure solutions or even freeze-dry them. A competent biochemist would never store his favourite protein at 37 °C in a highly reactive environment such as the intracellular milieu, where reactive oxygen species (for example hydroxyl radicals) are continually being generated that can destroy amino acids in proteins^{17,18}. It is also an environment where reactive sugars abound that can glycate proteins, where many enzymes are found that modify or destroy proteins (for example proteases), and where there are fatty acids that can function as detergents, and partially folded nascent chains that act as nuclei for aggregation. These protein-modifying processes must continually cause stochastic damage to well-formed cellular proteins, triggering their degradation². Even in pure solutions, multiple changes in primary sequence (deamidation of glutamines, and residue isomerization) occur over time (termed 'protein ageing'), leading to rapid degradation by proteasomes¹⁹. Although it generally functions as part of the ubiquitin–proteasome pathway, the 26S proteasome has the capacity to degrade certain unfolded or damaged proteins, including 'aged' or denatured proteins, without initial marking them by ubiquitylation. The frequency of such modifications and denaturing events must be high, because the free energy for stabilizing proteins is not great. Furthermore, in certain environments, the frequency of damage to cell proteins increases — for example after heat shock, in which many proteins are unfolded³, or during oxidative stress or inflammation, when oxygen radicals are generated in increased amounts¹⁷.

Box 1

Intracellular conditions that damage cell proteins

Temperature of 37 °C or higher (denaturing conditions).
Many reactive small molecules — these cause oxidation, deamidation, glycation or nitrosylation.
Many enzymes that modify proteins, for example proteases or kinases.
High salt concentrations (which favour dissociation of multimers)
Many fatty acids, which act like detergents.
Other unfolded proteins — nascent polypeptides, damaged or mutant polypeptides and insoluble inclusions are sticky.

Conclusion: to maintain a protein's function, avoid the intracellular milieu.

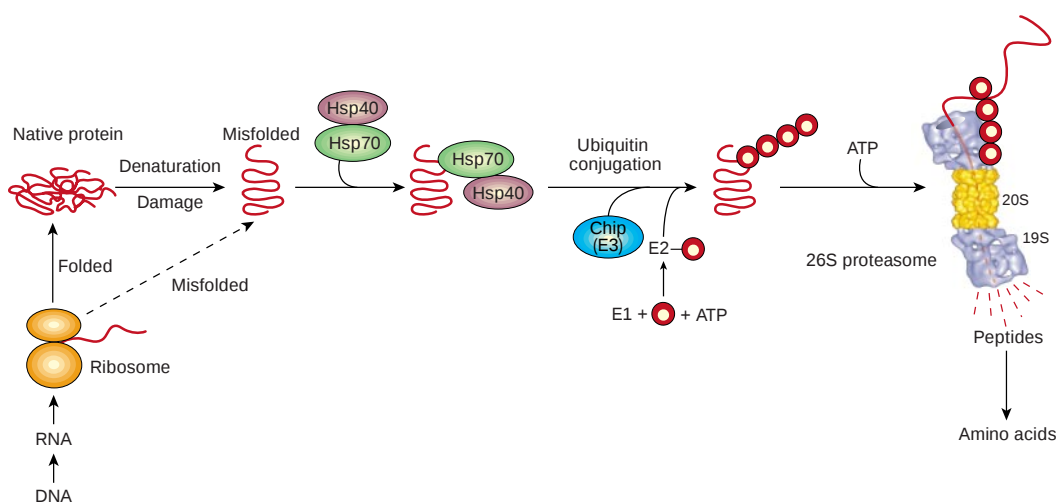


Figure 1 The ubiquitin–proteasome pathway. Molecular chaperones may function in protein folding and in the degradation of misfolded species. By associating with exposed hydrophobic domains, chaperones Hsp70/40 promote the folding of newly

synthesized proteins and favours their refolding. Alternatively, they can facilitate the recognition of abnormal proteins, leading to their ubiquitylation by CHIP, the E3, and their degradation by the 26S proteasome. The red circles represent ubiquitin.

Although many proteins are short-lived, the bulk of cell proteins are 'long-lived', and 1–2% of them are degraded per hour in mammals¹⁶. Even though these rates are physiologically regulated, the degradation of these long-lived proteins follows exponential decay kinetics, suggesting that a stochastic event triggers their destruction. Thus, for many — and perhaps most — proteins, a random event, spontaneous denaturation or chemical modification leading to unfolding might be the critical event triggering degradation².

Abnormal proteins may aggregate before degradation

A characteristic feature of denatured proteins is that they aggregate and leave solution (because of the tendency of normally buried hydrophobic domains to associate with one another). Similarly, when the production of misfolded proteins exceeds the cell's degradative capacity, these polypeptides often form intracellular aggregates before their rapid degradation^{3,7,20}. Indeed, the finding that many cloned, misfolded proteins also aggregate in the cytosol of bacteria has sometimes proved to be of major benefit to researchers, allowing the rapid isolation of the foreign polypeptides. The formation of such inclusions occurs when the cell's proteolytic systems and molecular chaperones, which normally prevent aggregation (for example Hsp70/40)^{13,14} or resolubilize microaggregates (Hsp104)²¹, cannot keep up with the rate of production of unfolded molecules.

Similar inclusions are found in various inherited and neurodegenerative diseases, in which these inclusions are ubiquitinated and associated with proteasomes³ (see also the article in this issue by Selkoe, p. 891), strongly suggesting a failure of the cell's degradative machinery³. When mammalian cells are treated with proteasome inhibitors, such inclusions appear rapidly, thus indicating that many substrates of the proteasome in normal cells are misfolded proteins³. As abnormal proteins accumulate, larger inclusions are found near the centrosome as a result of an active microtubule-dependent process that isolates the unfolded proteins in amorphous structures, which are often termed 'aggresomes'²². If the production of such abnormal proteins ceases and degradation is allowed to proceed, all cells, from bacteria²⁰ to neurons²³, can eliminate the proteinaceous inclusions. The mechanisms in mammalian cells for solubilizing and digesting these proteins are unknown, and further understanding of this process could have therapeutic applications for various protein-folding diseases.

The ubiquitin–proteasome pathway

A fundamental feature of protein degradation in all cells, and even mitochondria, is that it requires metabolic energy. This requirement for ATP is unexpected on thermodynamic grounds. Consequently, it was initially viewed as evidence for an ATP requirement for the function of lysosomes, then believed to be the only site of protein breakdown in cells. However, the selective degradation of abnormal proteins also requires ATP in cells that lack lysosomes (for example, bacteria)^{4,6}. This insight led to the development of cell-free preparations that catalyse the ATP-dependent degradation of abnormal proteins⁵. Attempts to understand this mysterious role for ATP led to the discovery of the involvement of ubiquitin conjugation in marking proteins for degradation²⁴ and of the 26S proteasome^{25,26}. In eukaryotes, ATP is initially essential to activate ubiquitin, which is then transferred to one of the cell's 20–40 different ubiquitin-carrier proteins (E2s). The exquisite selectivity of this pathway resides in the ubiquitin ligases, or E3s, which are specific for different protein substrates¹. Mammalian cells contain hundreds of different ubiquitin ligases, which, together with a specific E2, catalyse the formation of the ubiquitin chain on a limited number of protein substrates, triggering their rapid degradation by the 26S proteasome²⁷ (see below and Fig. 1).

Which E2s and E3s are involved in this cellular quality-control system and how they recognize misfolded proteins are still open questions. Several E3s are essential for the degradation of abnormal proteins in the secretory pathway (see the progress in this issue by Sitia and Braakman p. 891, and ref. 10), and certain E2s (Ubc4 and Ubc5) are necessary for the rapid degradation of cytosolic abnormal proteins and are induced as part of the heat-shock response¹¹. This transcriptional response is triggered in all cells by the appearance of unfolded proteins (for example, after exposure to increased temperatures or 'oxidants') and leads to an enhanced cellular content of chaperones and increased degradative capacity^{3,11,28}.

Recently, one E3, CHIP (for carboxy terminus of Hsp70-interacting protein)^{29,30}, has been identified in mammals that seems to ubiquitylate certain mutant proteins selectively in a process requiring the molecular chaperones Hsp70 or Hsp90, which bind to unfolded or hydrophobic domains^{13,14} and seem to facilitate substrate recognition by CHIP^{29,30}. These chaperones are normally among the most abundant cell proteins, and their levels rise further in harsh conditions that damage cell proteins^{3,11,28}. By binding to unfolded domains

of nascent proteins, they prevent their aggregation and facilitate normal folding in stressed cells, and promote the refolding of denatured molecules^{13,14}. The discovery that these chaperones also have a function in the selective degradation of abnormal proteins has important implications. Chaperone involvement in proteolysis, as well as folding, would provide an efficient quality-control mechanism, because unsuccessful folding would mean prolonged association of the substrate with the chaperones, leading to rapid degradation^{29–32}. However, it remains unclear whether CHIP catalyses the ubiquitylation of abnormal proteins generally or whether many E3s function in the degradation of different types of aberrant protein. It is also unclear whether ubiquitylation is essential for the proteasomal elimination of all abnormal proteins *in vivo*, because this process occurs efficiently in bacteria and archaea, which lack ubiquitin. In addition, *in vitro* eukaryotic 26S proteasomes degrade many unfolded proteins (for example, 'aged' calmodulin) without ubiquitin conjugation¹⁹.

ATP-dependent proteolytic machines

A very different and more general explanation of the requirement of ATP for the degradation of abnormal proteins emerged from our studies of this process in *E. coli* and mitochondria³³. Large proteolytic complexes were discovered whose activity was coupled to ATP hydrolysis^{33–35}. For example, the Lon (La) protease is a 600 kDa ATPase complex that rapidly degrades most abnormal proteins in *E. coli*^{33–35}. The Lon homologue in mitochondria is necessary for the viability of eukaryotes, presumably because it helps to prevent the build-up of aberrant proteins within this organelle^{36,37}. In addition, bacteria and mitochondria contain several other proteolytic complexes (namely ClpAP, ClpXP, HslUV and FtsH)^{5,33–36} that are composed of distinct ATPase and proteolytic subcomplexes. Like the proteasome, these enzymes are large complexes (at least 20–30-fold larger than typical proteases) that processively degrade proteins to oligopeptides. In each case, ATP hydrolysis appears necessary for the unfolding of substrates and for their delivery into a proteolytic sub-compartment, where the substrates are digested, thus avoiding the non-specific destruction of other cell constituents. These proteases are all heat-shock proteins^{3,5,10,33}, which means that they are coordinately induced together with molecular chaperones whenever cells generate large amounts of unfolded proteins (Table 1)^{3,10,28}. For example, they are often induced in bacteria that express cloned foreign proteins¹¹.

In all cells and mitochondria, these proteases function together with molecular chaperones in protein degradation³; for example, the rapid hydrolysis of certain abnormal proteins in *E. coli* requires the homologues of chaperones Hsp70/40 (ref. 32), whereas the degradation of others requires GroEL/ES chaperonins (also called Hsp60/10) (ref. 38). These large particles create an isolation chamber within which proteins can fold^{13,14}, yet they also seem to facilitate the proteolytic digestion of proteins that are hard to degrade³⁸.

How the cell's two major defence systems, molecular chaperones and proteasomes, function together in the elimination of abnormal proteins is not clear. Prolonged chaperone binding can help in the identification of proteins with unfolded domains^{30,32}, but the chaperones might also function as cofactors in maintaining the substrate in a soluble, unfolded state that facilitates proteolytic attack^{31,38}. An attractive ('triage') model is that a failure of the chaperones to perform protein folding or the refolding of damaged proteins leads to the recruitment of proteases or ubiquitylation enzymes to eliminate the potentially dangerous polypeptides^{3,30–32}.

The 26S proteasome

Whereas proteolysis in bacteria and mitochondria involves several specialized ATP-hydrolysing proteases, eukaryotic cells in their cytosol and nucleus contain only one such enzyme of enormous size (namely 50 subunits totalling 2.4 MDa) that degrades various types of substrate fed to it by multiple ubiquitin ligases. Understanding of the functioning of this proteolytic complex has advanced markedly

in recent years²⁷. The 26S proteasome is composed of a core 20S particle, in which proteins are digested to short peptides, and one or two 19S regulatory particles, responsible for substrate recognition and transport into the core particle²⁷. The 20S particle is composed of four stacked rings, whose subunits surround a central cavity. Its two inner β -rings form a central chamber containing the proteolytic sites, which face the central cavity. This organization ensures that protein digestion is isolated from the surrounding cytosol. Eukaryotic proteasomes contain two sites that cleave after hydrophobic residues, two after acidic residues and two after basic residues; thus they can cut most types of peptide bond^{5,27}. Substrates can enter the 20S particle only through a gated channel in the centre of the α -ring³⁹, which is normally maintained in a closed state, and access is controlled by the associated ATPases in the 19S particle⁴⁰.

The entry channel is quite narrow, and even in its open state it allows entry of only unfolded proteins^{39,40}. While excluding normal globular proteins, this architecture requires complex ATP-dependent mechanisms to recognize, unfold and linearize substrates and to inject them into the 20S proteasome⁴⁰. It is still unclear where in the 19S is located the initial binding site for ubiquitin chain and the enzymatic machinery for disassembling these chains^{1,27}. The base of the 19S contains a ring of six ATPases that provide added selectivity and somehow unfold proteins, translocate them and trigger gate-opening into the 20S particle⁴⁰. This elegant mechanism uses the energy in a considerable number of ATP molecules in degrading a protein, perhaps one-third of the ATP used by the ribosome in their synthesis⁴⁰. The essential features of this process emerged early in evolution; the proteasomes of Archaea function together with a very similar ATPase ring (termed PAN) that selects, unfolds and translocates substrates^{5,40}. With the emergence of eukaryotes, this process became linked to ubiquitin conjugation, which, like the architecture of the proteasome and its linkage to ATP hydrolysis, should be viewed as mechanisms to ensure that only unwanted proteins are selectively degraded.

Key questions and future prospects

The selective degradation of abnormal proteins has been known for about 30 years, but many fundamental questions remain unanswered. Although many adverse events during gene expression and protein folding, and many untoward postsynthetic events, trigger degradation, it is unclear to what extent this quality-control process contributes to overall protein turnover in normal cells. In addition, the mechanisms by which various types of misfolded polypeptides are recognized remain a mystery that has received only limited study. One important clue is that their degradation in all cells requires molecular chaperones, which probably function in substrate recognition^{3,30,32} but perhaps also in maintaining substrates in a soluble, easily digestible state^{31,38}.

This lack of mechanistic understanding is surprising and regrettable, because this process is of fundamental biological importance. The genetic deletion of virtually any subunit of the 26S proteasome in yeast prevents viability²⁷, and various point mutations leave the cells viable but very sensitive to conditions that increase the production of abnormal proteins (such as heat shock)²⁷. When unfolded proteins build up and exceed the cell's degradative capacity (for example, in the presence of proteasome inhibitors), cells activate the heat-shock response^{41,42} and induce the synthesis of more proteasomes⁴³. The continued accumulation of such proteins eventually triggers the activation of Jnk kinases and apoptosis^{3,44}. It is now well established that damage to cell proteins (just like damage to DNA or chromosomal organization) can activate the cell-death programme. However, the ability of unfolded proteins to activate apoptosis has received little attention, yet it probably has an important function in diseases where unfolded proteins accumulate and are associated with neurodegeneration. Importantly, certain cancers, especially multiple myeloma, are particularly sensitive to apoptosis induced by proteasome inhibitors, and one such inhibitor (Velcade) has recently been

approved for treatment of myeloma^{42,45}. These cancer cells generate very large amounts of abnormal immunoglobins and consequently are probably the cells that are most dependent on proteasome function for the continual elimination of abnormal proteins.

Evolution has provided multiple levels of protection against abnormal proteins — from proteasomes and chaperones to induction of the heat-shock genes and apoptosis. To what extent these mechanisms protect us against various human diseases is still uncertain. It is clear that in many diseases these proteolytic systems fail to prevent the accumulation of the damaging proteins, perhaps because of defects in this degradative machinery. Greater understanding of this quality-control process is therefore not only of scientific interest but might lead to new therapies. □

doi:10.1038/nature02263

- Glickman, M. H. & Ciechanover, A. The ubiquitin–proteasome proteolytic pathway: Destruction for the sake of construction. *Physiol. Rev.* **82**, 373–428 (2002).
- Goldberg, A. L. & Dice, J. F. Intracellular protein degradation in mammalian and bacterial cells. *Annu. Rev. Biochem.* **43**, 835–869 (1974).
- Sherman, M. & Goldberg, A. L. Cellular defenses against unfolded proteins: a cell biologist thinks about neurodegenerative diseases. *Neuron* **29**, 15–32 (2001).
- Goldberg, A. L. Degradation of abnormal proteins in *E. coli*. *Proc. Natl Acad. Sci. USA* **69**, 422–426 (1972).
- Zwickl, P., Goldberg, A. L. & Baumeister, W. in *Proteasomes: The World of Regulatory Proteolysis* (eds Wolf, D. H. & Hilt, W.) 8–20 (Landes Bioscience, Georgetown, Texas, 2000).
- Etlinger, J. & Goldberg, A. L. A soluble, ATP-dependent proteolytic system responsible for the degradation of abnormal proteins in reticulocytes. *Proc. Natl Acad. Sci. USA* **74**, 54–58 (1977).
- Klemes, Y., Etlinger, J. D. & Goldberg, A. L. Properties of proteins degraded rapidly in reticulocytes: intracellular aggregation of the globin molecules prior to hydrolysis. *J. Biol. Chem.* **256**, 8436–8444 (1981).
- Bunn, H. F. *et al.* *Hemoglobin: Molecular, Genetic, and Clinical Aspects* (Saunders, Philadelphia, 1986).
- Goldberg, A. L. & Goff, S. A. in *Maximizing Gene Expression* Ch. 9 (eds Reznikoff, W. & Gold, L.) 287–314 (Butterworths, Stoneham, Massachusetts, 1986).
- Kostova, Z. & Wolf, D. H. For whom the bell tolls: protein quality control of the endoplasmic reticulum and the ubiquitin–proteasome connection. *EMBO J.* **22**, 2309–2317 (2003).
- Goff, S. A., Voellmy, R. & Goldberg, A. L. in *Ubiquitin* Ch. 8 (ed. Rechsteiner, M.) 207–238 (Plenum, New York, 1988).
- Roche, E. & Sauer, R. T. SsrA-mediated peptide tagging caused by rare codons and tRNA scarcity. *EMBO J.* **18**, 4579–4589 (1999).
- Frydman, J. Folding of newly translated proteins *in vivo*: the role of molecular chaperones. *Annu. Rev. Biochem.* **70**, 603–647 (2001).
- Hartl, F. U. & Mayer-Hartl, M. Molecular chaperones in the cytosol: from nascent chain to folded protein. *Science* **295**, 1852–1858 (2002).
- Schubert, U. *et al.* Rapid degradation of a large fraction of newly synthesized proteins by proteasomes. *Nature* **44**, 770–774 (2000).
- Gronostajski, R., Pardee, A. B. & Goldberg, A. L. The ATP-dependence of the degradation of short- and long-lived proteins in growing fibroblasts. *J. Biol. Chem.* **260**, 3344–3349 (1985).
- Berlett, B. S. & Stadtman, E. R. Protein oxidation in aging, disease, and oxidative stress. *J. Biol. Chem.* **272**, 20313–20316 (1997).
- Grune, T., Reinheckel, T., Davies, K. J. Degradation of oxidized proteins in mammalian cells. *FASEB J.* **11**, 536–534 (1997).
- Tarcsa, E., Szymanska, G., Lecker, S., O'Connor, C. M. & Goldberg, A. L. Ca²⁺-free calmodulin and calmodulin damaged by *in vitro* aging are selectively degraded by 26S proteasomes without ubiquitylation. *J. Biol. Chem.* **275**, 20295–20301 (2000).
- Prouty, W. F. & Goldberg, A. L. Fate of abnormal proteins in *E. coli*: accumulation in intracellular granules before catabolism. *Nature New Biol.* **240**, 147–150 (1972).
- Glover, J. R. & Lindquist, S. Hsp104, Hsp70, and Hsp40: A novel chaperone system that rescues previously aggregated proteins. *Cell* **94**, 73–82 (1998).
- Johnston, J. A., Ward, C. L. & Kopito, R. R. Aggresomes: a cellular response to misfolded proteins. *J. Cell Biol.* **143**, 1883–1898.
- Yamamoto, A., Lucas, J. J. & Hen, R. Reversal of neuropathology and motor dysfunction in a conditional model of Huntington's disease. *Cell* **101**, 57–66 (2000).
- Ciechanover, A., Heller, H., Elias, S., Haas, A. L. & Herskho A. ATP-dependent conjugation of reticulocyte proteins with the polypeptide required for protein degradation. *Proc. Natl Acad. Sci. USA* **77**, 1365–1368 (1980).
- Hough, R., Pratt, G. & Rechsteiner, M. Purification of two high molecular weight proteases from rabbit reticulocyte lysate. *J. Biol. Chem.* **262**, 8303–8313 (1987).
- Waxman, L., Fagan, J. M. & Goldberg, A. L. Demonstration of two distinct high molecular weight proteases in rabbit reticulocytes, one of which degrades ubiquitin conjugates. *J. Biol. Chem.* **262**, 2451–2457 (1987).
- Voges, D., Zwickl, P., Baumeister, W. The 26S proteasome: a molecular machine designed for controlled proteolysis. *Annu. Rev. Biochem.* **68**, 1015–1068 (2000).
- Ananthan, J., Goldberg, A. L. & Voellmy, R. Abnormal proteins serve as eukaryotic stress signals and trigger the activation of heat-shock genes. *Science* **232**, 522–524 (1986).
- Meacham, G. C., Patterson, C., Zhang, W., Younger, J. M. & Cyr, D. M. The Hsc70 co-chaperone CHIP targets immature CFTR for proteasomal degradation. *Nature Cell Biol.* **3**, 100–105 (2001).
- Murata, S., Minami, Y., Minami, M., Chiba, T. & Tanaka, K. CHIP is a chaperone-dependent E3 ligase that ubiquitylates unfolded protein. *EMBO Rep.* **2**, 1133–1138 (2001).
- Wickner, S., Maurizi, M. & Gottesman, S. Posttranslational quality control: folding, refolding, and degrading proteins. *Science* **286**, 1888–1893 (1999).
- Huang, H. C., Sherman, M. Y., Kandror, O. & Goldberg, A. L. The molecular chaperone DnaJ is required for the degradation of a soluble abnormal protein in *E. coli*. *J. Biol. Chem.* **276**, 3920–3928 (2001).
- Goldberg, A. L. The mechanisms and functions of ATP-dependent proteases in bacterial and animal cells. *Eur. J. Biochem.* **203**, 9–23 (1992).
- Chung, C. H. Proteases in *Escherichia coli*. *Science* **262**, 372–374 (1993).
- Maurizi, M. R. Proteases and protein degradation in *Escherichia coli*. *Experientia* **48**, 178–201 (1992).
- Langer, T., Kaser, M., Klanner, C., Leonhard, K. AAA proteases of mitochondria: quality control of membrane proteins and regulatory functions during mitochondrial biogenesis. *Biochem. Soc. Trans.* **29**, 431–436 (2001).
- Suzuki, C. K., Rep, M., van Dijk, J. M., Suda, K., Grivell, L. A., Schatz, G. ATP-dependent proteases that also chaperone protein biogenesis. *Trends Biochem. Sci.* **22**, 118–123 (1997).
- Kandror, O., Sherman, M. Y. & Goldberg, A. L. Rapid degradation of an abnormal protein in *E. coli* proceeds through repeated cycles of association with GroEL. *J. Biol. Chem.* **274**, 37743–37749 (1999).
- Groll, M. *et al.* A gated channel into the proteasome core particle. *Nature Struct. Biol.* **7**, 1062–1067 (2000).
- Benaroudj, N., Zwickl, P., Seemüller, E., Baumeister, W. & Goldberg, A. L. ATP hydrolysis by the proteasome regulatory complex PAN serves multiple functions in protein degradation. *Mol. Cell* **11**, 69–78 (2003).
- Lee, D. H. & Goldberg, A. L. Proteasome inhibitors cause rapid induction of heat-shock proteins and trehalose, which together confer thermotolerance in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **18**, 30–38 (1998).
- Kisselev, A. F. & Goldberg, A. L. Proteasome inhibitors: from research tools to drug candidates. *Chem. Biol.* **8**, 739–758 (2001).
- Meiners, S. *et al.* Inhibition of proteasome activity induces concerted expression of proteasome genes and *de novo* formation of mammalian proteasomes. *J. Biol. Chem.* **278**, 21517–21525 (2003).
- Meriin, A. B. *et al.* Protein-damaging stresses activate c-Jun N-terminal kinase via inhibition of its dephosphorylation: a novel pathway controlled by HSP72. *Mol. Cell. Biol.* **19**, 2547–2555 (1999).
- Hideshima, T. *et al.* The proteasome inhibitor ps-341 inhibits growth, induces apoptosis, and overcomes drug resistance in human multiple myeloma cells. *Cancer Res.* **61**, 3071–3076 (2001).

Folding proteins in fatal ways

Dennis J. Selkoe

Center for Neurologic Diseases, Harvard Medical School, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA
(e-mail: dselkoe@rics.bwh.harvard.edu)

Human diseases characterized by insoluble extracellular deposits of proteins have been recognized for almost two centuries. Such amyloidoses were once thought to represent arcane secondary phenomena of questionable pathogenic significance. But it is now clear that many different proteins can misfold and form extracellular or intracellular aggregates that initiate profound cellular dysfunction. Particularly challenging examples of such disorders occur in the post-mitotic environment of the neuron and include Alzheimer's and Parkinson's diseases. Understanding some of the principles of protein folding has helped to explain how such diseases arise, with attendant therapeutic insights.

One of the most satisfying moments in scientific investigation occurs when previously disparate phenomena of unclear origin are shown to arise from a common principle. During the past decade, numerous aetiologically distinct diseases have been linked by the likelihood that they result from the progressive misfolding of specific proteins into aggregates that can injure and kill cells. Together, these disorders inflict enormous personal and societal burdens, making it crucial to understand their genesis and to learn how to prevent them.

The amyloidoses have traditionally been defined as diseases in which normally soluble proteins accumulate in the extracellular space of various tissues as insoluble deposits of 10 nm fibrils that are rich in β -sheet structure and have characteristic dye-binding properties^{1,2}. There are many examples of secreted, circulating proteins that can, under abnormal circumstances, be converted in part to highly stable extracellular fibrils. These include immunoglobulins in primary systemic amyloidosis and multiple myeloma, amylin in the diabetic pancreas, and small soluble proteins of uncertain function such as the amyloid β -peptide (A β) in Alzheimer's disease.

Although the specific polypeptides that comprise the deposits are different for each amyloidosis, the disorders have several key features in common. Perhaps foremost among them is the ability of proteins that are highly soluble in biological fluids to be converted gradually to insoluble filamentous polymers enriched in β -pleated sheet conformation. The common structural motif of virtually all amyloid fibrils consists of cross- β -sheets in which the peptide strands are arranged perpendicular to the long axis of the fibre. Although it was once thought that relatively few proteins have this propensity, recent data suggest that many soluble proteins can, under certain circumstances, undergo this conversion. Of particular significance is the finding that globular proteins with diverse sequences that are not currently associated with a protein-folding disease (for example, muscle myoglobin) can undergo aggregation *in vitro* into fibrils indistinguishable from those found in the amyloidoses^{3,4}. This finding supports the concept that aggregation into β -sheet-rich fibrils is a generic property of polypeptide chains regardless of sequence⁵.

Another general feature of protein-folding disorders is the prolonged period before clinical manifestations appear. Although the age of onset of symptoms varies widely among the different diseases and even among cases of one disease, most of these disorders become noticeable in middle or late life. There is a prolonged preclinical phase during which

proteins misfold, build up and progressively compromise cellular and tissue function. A portion of this long prodrome derives from the energetic barriers to the formation of misfolded species, including the fact that nucleation — the initial development of very small, metastable oligomers of a protein — is a kinetically unfavourable requirement for fibrillogenesis^{6,7}. It seems that time, rather than great age, is required, in that some aggressive protein-folding disorders can occur in young and early middle-aged individuals. In such cases, time still has a role but the fibrillogenic process requires less time overall because particular biochemical circumstances promote accelerated nucleation. One striking example is Down's syndrome, in which patients with trisomy 21 develop abundant A β aggregates in the brain as early as the age of ten, owing to lifelong overexpression of the β -amyloid precursor protein (APP), which is encoded on chromosome 21. Similarly, inherited missense mutations in amyloid-prone proteins can markedly accelerate their misfolding and fibrillogenesis, producing earlier disease onset than occurs with the wild-type isoform. For example, senile systemic amyloidosis arises late in life from the aggregation of wild-type transthyretin, whereas familial amyloidotic polyneuropathy I generally arises in mid-life from the accelerated aggregation of mutant transthyretin.

In this review, I describe briefly three types of amyloidosis: systemic, organ-limited and intracellular. I then examine how aberrant protein folding may occur and how misfolded proteins may disrupt the cell.

Systemic amyloidoses

The list of secreted, circulating proteins that are capable of producing extracellular amyloid deposits in multiple organs is long and growing (Table 1). Amyloidoses can arise when other pathological conditions cause a sharp increase in the concentration of an amyloid-prone polypeptide, as occurs for the serum amyloid A (SAA) protein during the acute-phase response accompanying inflammatory disorders such as rheumatoid arthritis, or chronic granulomatous infections such as tuberculosis. A 76-residue proteolytic fragment of wild-type SAA can then accumulate, misfold, aggregate and be deposited in the connective tissue of multiple organs, including spleen, kidney and liver. Multiple myeloma is associated with the overproduction by plasma cells of monoclonal immunoglobulins that accumulate and form deposits (as holoproteins and/or proteolytic fragments) in the extracellular space of various tissues. Primary systemic amyloidosis also involves progressive multi-tissue deposition of immunoglobulin light chains and their fragments. In some systemic amyloidoses, the basis for an

Table 1 Some members of the family of systemic extracellular amyloidoses

Clinical syndrome	Fibril subunit
Primary systemic amyloidosis	Intact immunoglobulin light chains or fragments thereof
Secondary systemic amyloidosis	Fragments of serum amyloid A protein
Familial Mediterranean fever	Fragments of serum amyloid A protein
Familial amyloidotic polyneuropathy 1	Mutant transthyretin and fragments thereof
Senile systemic amyloidosis	Wild-type transthyretin and fragments thereof
Familial amyloidotic polyneuropathy II	Fragments of apolipoprotein A-1
Haemodialysis-related amyloidosis	β_2 -Microglobulin
Finnish hereditary amyloidosis	Fragments of mutant gelsolin
Lysozyme amyloidosis	Full-length mutant lysozyme
Insulin-related amyloid	Full-length insulin
Fibrinogen α -chain amyloidosis	Fibrinogen α -chain variants

increased concentration of an amyloidogenic protein is a perturbation in its clearance. One striking example is haemodialysis-related amyloidosis, in which chronic haemodialysis can lead to progressive increase in β_2 -microglobulin in tissues because this amyloid-prone protein does not pass efficiently through the dialysis membrane and becomes highly concentrated in the post-dialysis serum of the patient.

In these and other systemic amyloid diseases, the precise reasons for selective tissue deposition and the mechanisms of cell dysfunction and organ failure are still under investigation (see below). Because systemic amyloid deposits can sometimes build up in very large amounts, it seems that they can cause injury in part by physical compression and microvascular compromise of adjacent tissue. However, as with the misfolded proteins that define certain neurodegenerative disorders (see below), the systemic extracellular amyloid fibrils and in particular their smaller, more diffusible precursor assemblies (oligomers and protofibrils) might also act in an amphipathic fashion to bind and perturb multiple cell-surface receptors and/or channels rather nonspecifically.

Organ-limited amyloidoses

The other broad class of amyloid disorders is that in which the fibrillogenic protein accumulates locally near its site of cellular production; that is, mostly in one organ (Table 2). Salient examples of this type include calcitonin deposition in medullary carcinoma of the thyroid, deposition of amylin in the pancreas in type II diabetes mellitus, deposition of atrial natriuretic factor in atrial amyloidosis of the heart, and A β deposition in Alzheimer's disease. Although these proteins enter the systemic circulation, their high local production in these conditions apparently leads to levels that exceed the critical concentrations for oligomerization and fibrillogenesis solely in that organ. In these disorders, the amyloid fibrils and their precursor aggregates are deposited in the extracellular space of the respective organs and somehow compromise local cell viability. However, there are also a growing number of diseases in which the accumulation of different misfolded proteins occurs intracellularly, particularly in the nervous system, as described in the next section.

Protein misfolding and age-related neurodegeneration

During the past two decades, there has been a marked transformation in our understanding of the causes and pathogenesis of age-related degenerations of the brain. Once fraught with mechanistic ignorance, this class of disorders has become recognized as likely to be mediated by the progressive accumulation of β -sheet-rich protein aggregates. A range of brain disorders of previously unknown cause now falls into this category (Table 3). The most common neurodegenerative disease, Alzheimer's, is the premier example of this. The second most common, Parkinson's, is increasingly considered to involve a protein-folding problem. Two special features of most of these disorders of the central nervous system are that the aggregates accumulate inside cells and that the ultrastructure of the aggregates is not the same as that of extracellular amyloid fibrils. This raises the

question of whether such disorders should be classified with the amyloidoses. But because the dye-binding properties (for example, birefringent labelling by Congo red), insolubility and X-ray diffraction pattern (high β -sheet content) of at least some of these deposits seem to be similar to those of classic amyloid fibrils, it is reasonable to consider them a special form of amyloidosis.

A wide variety of proteins of diverse sequence can form intraneuronal filamentous deposits. In Parkinson's disease, α -synuclein, a soluble cytoplasmic protein with a 'natively unfolded' structure, misfolds and accumulates in spherical filamentous masses (Lewy bodies) within selected neuronal cell bodies, particularly the dopaminergic and noradrenergic brainstem neurons whose premature death defines the disease biochemically. A recently recognized group of protein-folding diseases of the central nervous system are the polyglutamine repeat disorders, such as Huntington's disease and several forms of familial spinocerebellar ataxia⁸. Here, unstable DNA mutations in different genes result in the lengthening of glutamine stretches in the cognate proteins, which subsequently misfold and accumulate in the nuclei and cytoplasm of certain neurons. Although the appearance of microscopically detectable filamentous inclusions is variable and subtle compared with the robust formation of neurofibrillary tangles in Alzheimer's disease and Lewy bodies in Parkinson's disease, these highly stable, polyglutamine-rich proteins aggregate and gradually cause the dysfunction and death of their host neurons. It seems that the accumulation of misfolded proteins has particularly dire consequences in the post-mitotic milieu of the adult neuron.

Alzheimer's disease is virtually the only brain disorder that is defined by the accumulation of amyloid-forming proteins both extracellularly (A β) and intracellularly (tau). There has been lively debate about which of these, if either, has precedence in the pathogenic mechanism. The issue has been largely resolved by the finding that inherited mutations in APP, or in one of the proteases (called presenilin/ γ -secretase) that cleaves APP to release A β , cause aggressive, early-onset forms of Alzheimer's disease⁹. In contrast, inherited mutations in the tau protein do not produce Alzheimer's disease but cause the less common but equally devastating disorder, frontotemporal dementia with parkinsonism, in which tau-containing neurofibrillary tangles accumulate in the absence of extracellular amyloid¹⁰.

A remarkable example of protein misfolding causing neurodegeneration is that of the prion disorders. In humans, these include Creutzfeldt–Jacob disease and new-variant Creutzfeldt–Jacob disease, the latter being a human derivative of bovine spongiform encephalopathy ('mad cow disease'). These disorders are very closely related to a range of slowly progressive infectious diseases in certain lower mammals, including the defining disorder, scrapie, in sheep¹¹. Very rarely, the normal cellular prion protein (PrP) can undergo a change in its conformation that creates a metastable conformer, which can in turn bind and 'convert' another wild-type prion protein into this same pathological isoform. This misfolding process can apparently occur spontaneously at a very low frequency but occurs more readily if the protein bears certain missense mutations. Prion encephalopathies can thus be inherited, but they also occur in an infectious form: exposure of a healthy subject to small amounts of pathologically conformed but non-mutant prion protein (for example, through a corneal transplant from an individual with

Table 2 Some of the organ-limited extracellular amyloidoses

Clinical syndrome	Fibril subunit
Alzheimer's disease	Amyloid β -peptide
Spongiform encephalopathies	Full-length prion protein or fragments thereof
Hereditary cerebral haemorrhage with amyloidosis	Amyloid β -peptide or cystatin C
Type II diabetes	Amylin (islet amyloid polypeptide)
Medullary carcinoma of the thyroid	Procalcitonin
Atrial amyloidosis	Atrial natriuretic factor

Creutzfeldt–Jacob disease) can lead to progressive and fatal brain degeneration. Exactly how prion conversion and propagation of the disease occur is under study. One theory favours an ‘instructive’ dimerization of a scrapie-type prion protein conformer with a wild-type molecule that converts the latter to the former form; this can then be magnified geometrically to propagate the process. Alternatively, small oligomers of the scrapie-type conformer might serve as ‘seeds’ for the progressive addition and conformational conversion of wild-type monomers, in a mechanism akin to those postulated in systemic amyloidoses and Alzheimer’s disease. It should be emphasized that, like the prion protein, numerous amyloidogenic proteins can oligomerize and form toxic aggregates in their wild-type state, for example tau in Alzheimer’s disease and α -synuclein in idiopathic Parkinson’s disease.

Mechanism of protein misfolding

A principal unanswered question about all of these disorders is the precise manner in which natively soluble proteins of distinct primary structure undergo partial unfolding and aberrant refolding to produce highly stable oligomers and polymers with novel properties. It is clear that supraphysiological concentrations, coupled with prolonged time and certain biochemical conditions, underlie the initiation of the oligomerization process. One instructive example that has been studied in some detail is lysozyme, a protein that causes a fatal systemic amyloidosis only when it bears one of two inherited missense mutations, Ile 56→Thr or Asp 67→His. Crystallography of the wild-type and mutant variants revealed subtle but structurally significant changes at the interface between the α - and β -domains that suggest that this region is less constrained in both mutants¹². In agreement, circular-dichroism studies during conditions of heat denaturation that lead to fibril formation showed that both mutants are less thermostable. Moreover, near the midpoint of the heat-driven unfolding process, the mutants formed partly folded intermediates with considerable helical structure but no persistent tertiary interactions¹². Wild-type lysozyme unfolds into a partly structured intermediate only when thermal denaturation is performed at very low pH (ref. 13).

It seems that such partly folded forms, which have exposed hydrophobic regions and are therefore prone to self-aggregation, can exist in equilibrium with the native protein. In the case of the mutant lysozymes, the molten-globule-like intermediates have persistent structure in the α -domain but lack stable, native-type structure in the β -domain¹⁴. It has therefore been proposed that the transient, partly folded forms of the mutant lysozymes associate through their unstable β -domains. The emergence of stable β -structure through such intermolecular self-association might provide a template (seed) for the recruitment of additional peptide chains that ultimately form the hydrogen-bonded, mainly cross- β -sheet core structure of the insoluble amyloid fibril¹² (Fig. 1). In agreement with this model, the NMR structure of a domain of the prion protein (PrP(121–231)) indicates that amino acids that are mutated in inherited prion diseases are involved in the maintenance of the hydrophobic core¹⁵. Because microscopically visible fibril deposition is not an obligatory feature of the prion diseases, these principles are likely to apply also to pre-fibrillar, oligomeric assemblies of PrP and presumably other amyloidogenic proteins.

Table 3 Some human brain diseases characterized by progressive misfolding and aggregation of proteins

Disease	Protein	Locus
Alzheimer’s disease	Amyloid β -protein Tau	Extracellular plaques Tangles in neuronal cytoplasm
Frontotemporal dementia with parkinsonism	Tau	Tangles in neuronal cytoplasm
Parkinson’s disease; dementia with Lewy bodies	α -Synuclein	Neuronal cytoplasm
Creutzfeldt–Jakob disease; ‘mad cow disease’*	Prion protein (PrP ^{Sc})	Extracellular plaques Oligomers, inside and outside neurons
Polyglutamine expansion diseases (Huntington’s disease, spinocerebellar ataxias, and so on)*	Long glutamine stretches within certain proteins	Neuronal nuclei and cytoplasm
Amyotrophic lateral sclerosis*	Superoxide dismutase	Neuronal cytoplasm

*Figures reproduced from Greenfield’s *Neuropathology* (copyright Hodder Arnold)

How do misfolded proteins cause cell dysfunction?

Perhaps even more complex than the mechanisms by which proteins misfold are the mechanisms by which they inflict cellular injury. This question once focused primarily on the mature, highly stable amyloid fibrils, but now attention has shifted to the apparent cytotoxicity conferred by metastable intermediates of the fibrillogenic process. In diseases in which unmistakable amyloid formation occurs, deposits of insoluble fibrils are in equilibrium with smaller, more diffusible assemblies that are invisible by light microscopy but can be detected by biochemical methods such as gel electrophoresis and size-exclusion chromatography. In the growing number of diseases in which protein misfolding and oligomerization occur but no definite amyloid fibrils are seen, these smaller diffusible assemblies might be responsible in part for cytotoxicity. Limited studies of such species in human-tissue extracts indicate that they might include dimers, trimers, tetramers and other larger oligomers. But most of the information about the properties of such intermediates has been gleaned from biophysical (*in vitro*) and cell culture studies of synthetic or recombinant versions of the amyloidogenic proteins. The concern here is that perhaps only a small proportion of these artificial assemblies closely resemble the natural oligomers that occur *in vivo*.

Mouse models that transgenically overexpress the polypeptide precursors of amyloid subunits *in vivo* offer another approach. One closely studied example is that of mice overexpressing mutant human APP as a model of the amyloidosis of Alzheimer’s disease. In some such mouse lines, presynaptic nerve terminals and neuronal cell bodies are decreased about 30% compared with non-transgenic controls at the age of 2–4 months, when concentrations of soluble A β are increasing but well before A β deposition (that is, amyloid plaque formation) has begun¹⁶. Comparisons of transgenic lines with different APP expression levels suggest that these decreases in presynaptic terminals are dependent on A β concentrations, not on the A β plaque burden¹⁷. This animal work fits nicely with growing evidence that

memory and cognitive deficits in Alzheimer's disease and in its harbinger, minimal cognitive impairment, correlate better with cerebral A β concentrations than with amyloid plaque numbers¹⁸ and correlate best with the soluble pool of cerebral A β , which includes soluble oligomers^{19,20}. In the brains of subjects who have died (for other reasons) with only mild clinical impairment, the concentrations of soluble A β monomers and oligomers in the cerebral cortex correlate with the degree of synaptic loss²⁰.

Electrophysiological studies of young mice that are transgenic for human APP bearing Alzheimer's-disease-causing missense mutations reveal significant deficits in basal synaptic transmission and/or long-term potentiation (LTP, an electrophysiological correlate of synaptic plasticity) in the hippocampus, well before the development of microscopically detectable A β deposits (reviewed in ref. 21). However, the nature of the 'synaptotoxic' A β species in the brain is difficult to define, because these animals accumulate a mixture of A β forms (monomers, soluble oligomers, insoluble oligomers, and some insoluble amyloid fibrils) that are likely to exist in dynamic equilibrium. In certain cultured cells expressing mutant human APP, natural oligomers of human A β are formed soon after generation of the peptide within intracellular vesicles and are subsequently secreted from the cell at low nanomolar concentrations²². Cerebroventricular microinjection of cell medium containing these soluble oligomers and abundant monomers (but no amyloid fibrils) potently inhibits hippocampal LTP in adult rats. Pretreatment of the medium with a protease that selectively degrades A β monomers but not oligomers fails to prevent LTP inhibition. In contrast, treatment of the cells with an inhibitor of γ -secretase (one of the two proteases that generate A β from APP) markedly decreases oligomer formation at doses that still allow appreciable monomer production, and such medium no longer disrupts LTP²². These experiments allow one to attribute an inhibition of synaptic function *in vivo* specifically to soluble oligomers, not monomers or fibrils, of secreted A β .

In the polyglutamine expansion disorders, similar phenomena may be occurring, but intraneuronally. For example, in mouse models of spinocerebellar ataxia-1, in which the protein ataxin-1 bears long glutamine repeats and can form microscopically visible intranuclear inclusions in selected neurons, the most vulnerable neurons develop large filamentous inclusions only late in the course of neurodegeneration, whereas neurons showing early inclusion formation are more resistant to cell death²³. In this sense, inclusions of a misfolded protein might be protective because they sequester the aggregates, at least temporarily. In α -synuclein aggregation in Parkinson's disease, loss of dopaminergic neurons in a *Drosophila* model expressing human α -synuclein was prevented by co-expressing the heat shock protein Hsp70, one of numerous molecular chaperones that guide the correct folding of polypeptides²⁴. This work and similar approaches in the polyglutamine disorders^{25,26} indicate that such diseases are indeed disorders of protein folding, and they remind us that chaperones and other powerful compensatory mechanisms — such as activation of the ubiquitin–proteasome system — can decrease the accumulation of misfolded proteins or else enhance their clearance.

Whether early-assembly forms or mature polymers or both are important for dysfunction, the question of precisely how such misfolded species perturb cellular homeostasis must also be answered. Many potential mechanisms for the adverse effects of misfolded proteins have been proposed. In the example of Alzheimer's disease, the accumulation of diffusible A β oligomers could lead to their non-specific binding to receptors and channel proteins on the synaptic plasma membrane, thus interfering with numerous signal-transduction cascades. One consequence of exposing cultured neurons to synthetic A β assemblies is to alter calcium homeostasis and another is to enhance the generation of free radicals, with resultant oxidative damage to membrane lipids and proteins. In the polyglutamine diseases, it could be that small oligomers of the expanded protein adhere to and perturb intracellular membranes and/or sequester vital proteins

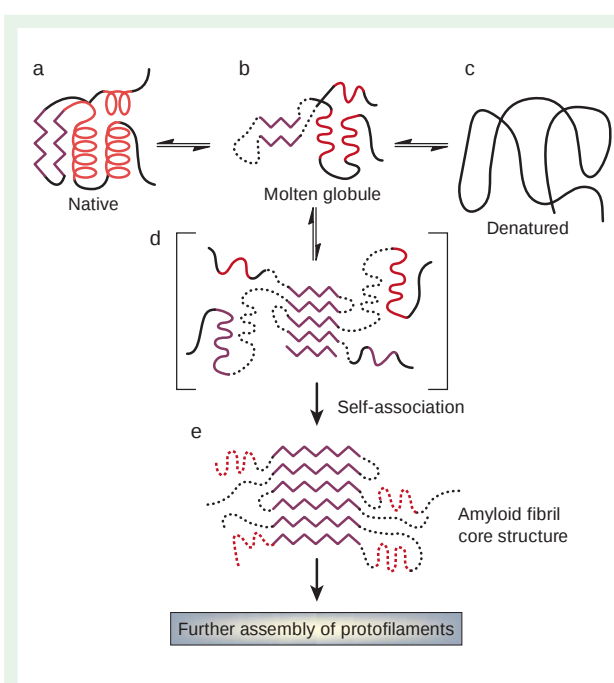


Figure 1 Proposed mechanism for lysozyme amyloid fibril formation. A partly folded, molten globule-like form of the protein (**b**), distinct from the native (**a**) and denatured (**c**) states, self-associates through the β -domain (**d**) to initiate oligomer formation. This provides a template for the further deposition of protein and for the development of the stable, mainly β -sheet, core structure of the protofilaments within an amyloid fibril (**e**). The undefined regions in **e** represent the likelihood that not all of the polypeptide sequence is involved in the cross- β structure. The nature of this residual structure in **e** is not known, and the figure is not intended to represent any defined secondary structural type¹². Purple, β -sheet structure; red, helical structure; dotted lines, undefined structure. (Modified from ref. 12.)

such as transcription factors. One can assume that the cell biological pathways will differ among the various amyloidogenic proteins, among the targeted cells and tissues, and even among the clinicopathological stages within any one amyloid disease.

Conclusion

The insidious accumulation of misfolded proteins has dire consequences for the organism. Further progress on the two key questions discussed above — how soluble proteins begin to misfold and how the resultant oligomers initiate cell dysfunction — will offer exciting prospects for specific molecular interventions. Some of the treatment modalities that will probably emerge, for example small-molecule inhibitors of monomer–monomer binding, might prove to be useful in treating or preventing more than one amyloidotic disease. We are also likely to see major advances in imaging the extracellular and perhaps even intracellular protein aggregates non-invasively, allowing therapy to be followed dynamically. In addition, the fact that even highly soluble globular proteins not currently implicated in human disease can, under the wrong biochemical circumstances, form filamentous polymers that confer cytotoxicity⁴ suggests that numerous other protein-folding disorders await recognition. □

doi:10.1038/nature02264

1. Glenner, G. G. Amyloid deposits and amyloidosis: the β -fibrilloses (first of two parts). *N. Engl. J. Med.* **302**, 1283–1292 (1980).
2. Glenner, G. G. Amyloid deposits and amyloidosis: the β -fibrilloses (second of two parts). *N. Engl. J. Med.* **302**, 1333–1343 (1980).
3. Fandrich, M., Fletcher, M. A. & Dobson, C. M. Amyloid fibrils from muscle myoglobin. *Nature* **410**, 165–166 (2001).
4. Bucciantini, M. *et al.* Inherent toxicity of aggregates implies a common mechanism for protein

- misfolding diseases. *Nature* **416**, 507–511 (2002).
5. Dobson, C. M. The structural basis of protein folding and its links with human disease. *Phil. Trans. R. Soc. Lond. B* **356**, 133–145 (2001).
6. Jarrett, J. T. & Lansbury, P. T. Seeding 'one-dimensional crystallization' of amyloid: A pathogenic mechanism in Alzheimer's disease and scrapie? *Cell* **73**, 1055–1058 (1993).
7. Perutz, M. F. & Windle, A. H. Cause of neural death in neurodegenerative diseases attributable to expansion of glutamine repeats. *Nature* **412**, 143–144 (2001).
8. Zoghbi, H. Y. & Orr, H. T. Glutamine repeats and neurodegeneration. *Annu. Rev. Neurosci.* **23**, 217–247 (2000).
9. Selkoe, D. J. & Podlisny, M. B. Deciphering the genetic basis of Alzheimer's disease. *Annu. Rev. Genomics Hum. Genet.* **3**, 67–99 (2002).
10. Hutton, M. Missense and splice site mutations in tau associated with FTDP-17: multiple pathogenic mechanisms. *Neurology* **56**, S21–S25 (2001).
11. Prusiner, S. B. Neurodegenerative diseases and prions. *N. Engl. J. Med.* **344**, 1516–1526 (2001).
12. Booth, D. R. *et al.* Instability, unfolding and aggregation of human lysozyme variants underlying amyloid fibrillogenesis. *Nature* **385**, 787–793 (1997).
13. Haezebrouck, P. *et al.* An equilibrium partially folded state of human lysozyme at low pH. *J. Mol. Biol.* **246**, 382–387 (1995).
14. Radford, S. E. & Dobson, C. M. Insights into protein folding using physical techniques: studies of lysozyme and α -lactalbumin. *Phil. Trans. R. Soc. Lond. B* **348**, 17–25 (1995).
15. Riek, R. *et al.* NMR structure of the mouse prion protein domain PrP(121–231). *Nature* **382**, 180–182 (1996).
16. Hsia, A. Y. *et al.* Plaque-independent disruption of neural circuits in Alzheimer's disease mouse models. *Proc. Natl Acad. Sci. USA* **96**, 3228–3233 (1999).
17. Mucke, L. *et al.* High-level neuronal expression of A β 1–42 in wild-type human amyloid protein precursor transgenic mice: synaptotoxicity without plaque formation. *J. Neurosci.* **20**, 4050–4058 (2000).
18. Naslund, J. *et al.* Correlation between elevated levels of amyloid β -peptide in the brain and cognitive decline. *J. Am. Med. Assoc.* **283**, 1571–1577 (2000).
19. Kuo, Y.-M. *et al.* Water-soluble A β (N-40, N-42) oligomers in normal and Alzheimer disease brains. *J. Biol. Chem.* **271**, 4077–4081 (1996).
20. Lue, L. F. *et al.* Soluble amyloid beta peptide concentration as a predictor of synaptic change in Alzheimer's disease. *Am. J. Pathol.* **155**, 853–862 (1999).
21. Selkoe, D. J. Alzheimer's disease is a synaptic failure. *Science* **298**, 789–791 (2002).
22. Walsh, D. *et al.* Naturally secreted oligomers of the Alzheimer amyloid β -protein potently inhibit hippocampal long-term potentiation *in vivo*. *Nature* **416**, 535–539 (2002).
23. Watase, K. *et al.* A long CAG repeat in the mouse Sca1 locus replicates SCA1 features and reveals the impact of protein solubility on selective neurodegeneration. *Neuron* **34**, 905–919 (2002).
24. Auluck, P. K., Chan, H. Y., Trojanowski, J. Q., Lee, V. M. & Bonini, N. M. Chaperone suppression of alpha-synuclein toxicity in a *Drosophila* model for Parkinson's disease. *Science* **295**, 865–868 (2002).
25. Warrick, J. M. *et al.* Suppression of polyglutamine-mediated neurodegeneration in *Drosophila* by the molecular chaperone HSP70. *Nature Genet.* **23**, 425–428 (1999).
26. Opal, P. & Zoghbi, H. Y. The role of chaperones in polyglutamine disease. *Trends Mol. Med.* **8**, 232–236 (2002).

Therapeutic approaches to protein-misfolding diseases

Fred E. Cohen¹ & Jeffery W. Kelly²

¹University of California at San Francisco, Department of Cellular and Molecular Pharmacology, Genentech Hall, 600 16th Street N472J, San Francisco, California 94107, USA (e-mail: cohen@cmpharm.ucsf.edu)

²Department of Chemistry and the Skaggs Institute of Chemical Biology, The Scripps Research Institute, 10550 North Torrey Pines Rd, BCC 265, La Jolla, California 92037, USA (e-mail: jkelly@scripps.edu)

Several sporadic and genetic diseases are caused by protein misfolding. These include cystic fibrosis and other devastating diseases of childhood as well as Alzheimer's, Parkinson's and other debilitating maladies of the elderly. A unified view of the molecular and cellular pathogenesis of these conditions has led to the search for chemical chaperones that can slow, arrest or revert disease progression. Molecules are now emerging that link our biophysical insights with our therapeutic aspirations.

How and whether a protein folds *in vivo* are influenced primarily by its amino-acid sequence and the cellular environment surrounding the polypeptide chain¹. Protein folding in the cell occurs either in the cytoplasm or within the secretory pathway (see the review in this issue by Dobson, page 884). ATP-dependent chaperones in the cytoplasm and the lumen of the endoplasmic reticulum (ER) collaborate to give a polypeptide chain several opportunities to fold. If folding is unsuccessful, the polypeptide is directed to the proteasome — a large multisubunit protease — for degradation (see the review in this issue by Goldberg, page 895). Ensuring accuracy in protein folding is crucial for maintaining proper cellular function.

Several diseases are caused by a loss of protein function because the misfolded proteins have been degraded by the proteasome. Cystic fibrosis and α 1-antitrypsin (α 1-AT) deficiency are well-known heritable diseases that fall into this category^{2–4}. Other misfolding diseases are caused by conformational changes coupled to the aggregation of misfolded proteins outside the cell, beyond the influence of intracellular quality-control systems. The major representatives of these disorders are the amyloidoses, in which protein aggregation in the extracellular space is associated with the appearance of a toxic function^{5,6}.

Proteins that use the secretory pathway for folding and trafficking are over-represented in protein-misfolding diseases. These proteins fold in the ER and then proceed through the secretory pathway to their final destination, where the environment can be very different. For example, lysosomal enzymes have to fold at neutral pH to leave the ER, yet their sequences are optimized for folding and stability at pH 5 (lysosomal pH). So mutations in lysosomal proteins can adversely alter folding in the ER, leading to degradation by the quality-control machinery. Ironically, many of the variants that cause heritable lysosomal storage diseases would be able to fold into functional enzymes in the lysosome, if only they could get there^{7,8}. In contrast, amyloidogenic proteins seem to fold efficiently in the ER but get into trouble in their destination environments, where they misfold and misassemble, a process known as amyloidogenesis. The thermodynamic principles behind protein folding are discussed in Box 1.

Therapeutic intervention can take the form of drugs that reduce the production of a certain protein, drugs that increase the rate of clearance of misfolded/aggregated pro-

teins, and drugs that increase the native-state stability or increase the kinetic barrier to misfolding aggregation (see Box 1).

In this review, we focus on small molecules, antibodies and dominant-negative protein subunits that increase the stability of the native state and/or increase the activation barrier associated with misfolding because these approaches are innovative and likely to be efficacious.

Small-molecule chaperones and treatment of disease

Many wild-type (WT) proteins fold inefficiently in the ER, even with the aid of the heat shock protein 40/70/90 (hsp40, hsp70 and hsp90) chaperone system that facilitates the refolding or retrotranslocation of misfolded proteins back into the cytoplasm, where they are degraded by the proteasome. Mutations further reduce folding efficiency by altering the energetics of proper folding. However, small-molecule ligands that bind to the native state of these mutant proteins can stabilize the native state or destabilize the transition state (see Box 1) to compensate for the influence of the mutation (Fig. 1).

Bouvier and colleagues⁹ have shown that selective V2-receptor antagonists (for example, SR121463A and VPA985) can lead to a 15-fold increase in expression of functional V2 vasopressin receptor at the cell surface, rescuing these receptors from proteasome degradation. Antagonists that were not cell permeable were ineffective, showing the importance of intercepting the misfolded protein in the cell and most likely in the ER. Chemical chaperones are small molecules that bind to a protein, stabilize the folded state and thereby reduce protein misfolding. The cell-permeable chemical chaperones were able to functionally complement more than six V2 receptor variants, some of which are associated with nephrogenic diabetes insipidus. These investigators also used chemical chaperones to complement the misfolding tendencies of variants of the δ -opioid receptor¹⁰.

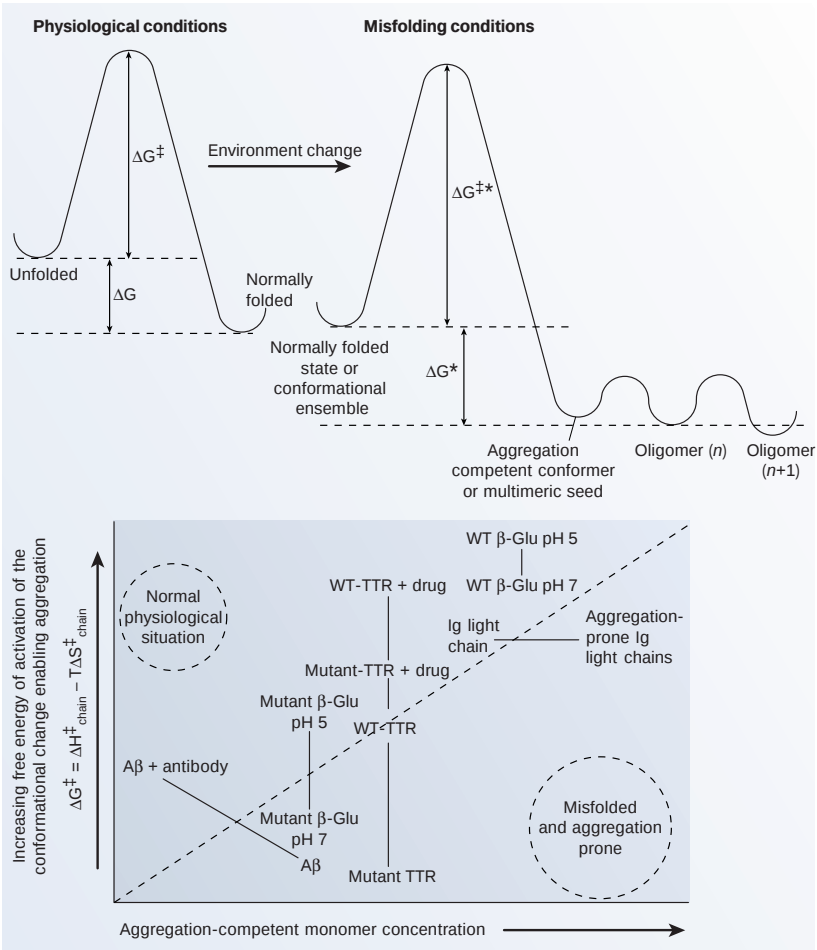
Conn and colleagues¹¹ have demonstrated that it is possible to use indoles, quinolones and erythromycin macrolides to complement the secretion-deficient loss-of-function variants of the gonadotropin-releasing hormone receptor (GnRHR) in transiently transfected COS-7 cells. ER-permeable agonists and antagonists with a high affinity for WT GnRHR were most efficient at chaperoning several, but not all, GnRHR variants out of the ER, and enabling their trafficking to the plasma membrane where they function. In an analogous approach, defects in rhodopsin folding in the

Box 1

The thermodynamics of protein folding

In discussing therapeutic approaches to diseases of protein misfolding, it is useful to consider the kinetic and thermodynamic underpinnings (see figure). Most proteins easily find their native state because the free energy of activation ($\Delta G^\ddagger$, which dictates rate) separating the unfolded and transition states is surmountable, and the folded state is substantially more stable than the unfolded state (the difference in free energy, ΔG , dictates the relative populations of folded and unfolded proteins). For aggregation-prone proteins, it has become clear that the misfolded oligomeric state can be more stable than the functional three-dimensional structure under certain conditions, including those associated with a high-denaturation stress^{47,48}. Fortunately, the kinetic barrier, $\Delta G^\ddagger$, separating the native and misfolded states is usually insurmountable on a biologically relevant time frame, and, in most cases, intrinsic clearance mechanisms operate more efficiently than the misfolding process^{19,49}. Because the misfolding process involves multimerization of individual chains, $\Delta G^\ddagger$ has two components, one relating to the conformation of the monomer and a second relating to multimerization that is dependent on concentration. The asterisk indicates that this $\Delta G^\ddagger$ is the barrier between the monomeric conformer and a multimer that can oligomerize efficiently.

The free-energy diagram in the top left panel of the figure indicates to what extent a given sequence under native conditions prefers a normally folded conformation. The difference in free energy (ΔG) between the native state and the unfolded ensemble of conformations (unfolded state) dictates the relative population of each state. The free energy of activation ($\Delta G^\ddagger$) governs how fast one conformation will convert to another. Under denaturing conditions, a misfolded aggregation-competent conformation can be preferred to the normally folded native state as a consequence of the interplay between sequence and environment¹, as shown in the top right panel of the figure. For example, the environment surrounding the polypeptide can change as a result of intracellular or extracellular alterations. ΔG and $\Delta G^\ddagger$ have implications analogous to those described above. The bottom panel of the figure illustrates the interplay between sequence, monomer concentration and the presence of drug or



antibody in altering the free energy of activation, $\Delta G^\ddagger = \Delta H^\ddagger - T\Delta S^\ddagger$, for amyloidogenesis. Situations below the diagonal permit amyloidogenesis, whereas those found above the diagonal resist misassembly.

After this energetic analysis, it is possible to consider therapeutics in three classes: interventions that decrease production; small molecules, antibodies or dominant-negative protein subunits that increase native state stability and/or the kinetic barrier to misfolding/aggregation; and approaches that increase clearance of the misfolded monomer, soluble multimer or insoluble amyloid.

disease retinitis pigmentosa were corrected by using retinal-based ligands¹², as were human *ether-a-gogo-related* gene (*HERG*) mutations by channel blockers¹³.

The most prevalent lysosomal storage disease, Gaucher's disease, is caused by the inefficient folding and trafficking of certain variants of β -glucocerebrosidase (β -Glu). The most prominent mutation (N370S) and other β -Glu mutations were thought to alter active-site residues. However, it is now becoming clear that several of these mutations interfere with folding in the ER and/or trafficking. The intracellular activity of misfolded variants such as N370S can be partly restored by administering cell-permeable β -Glu inhibitors that serve as chemical chaperones. These inhibitors seem to provide a template that complements the active site of β -Glu, stabilizing the native state in the ER, allowing β -Glu to leave the ER and be trafficked to the lysosome. The high concentration of the glycolipid substrate in the

lysosomes of β -Glu-deficient cells can competitively displace the inhibitor, allowing the enzyme to function. Kelly and colleagues⁸ have shown that a subset of N-alkylated deoxynojirimycins and simpler six-membered ring N-heterocycles increase the activity of β -Glu within human cells⁸; the best compounds discovered so far double β -Glu activity. This should be sufficient to ameliorate the pathology of Gaucher's disease and could offer a convenient alternative to intravenous enzyme replacement therapy; however, relevant transgenic animal models to test this approach are not yet available.

Fan *et al.*¹⁴ have shown that it is possible to use the small molecule 1-deoxy-galactonojirimycin, a potent competitive inhibitor of α -galactosidase A (α -Gal A), to rescue the misfolding and mistrafficking associated with certain Fabry-disease-causing mutations. These chemical chaperones facilitate the proper folding and trafficking of α -Gal A variants to the lysosome, increasing enzyme activity

seven- to eightfold in cells when used at sub-inhibitory intracellular concentrations. Biophysical experiments suggest that these chemical chaperones stabilize the enzyme. Oral administration of 1-deoxygalactonojirimycin to transgenic mice with Fabry disease expressing a variant of α -Gal A increased the enzyme activity significantly in several organs¹⁴. Chaperoning of α -Gal A can also be accomplished by galactose infusions in humans¹⁵.

Other loss-of-function diseases, including cystic fibrosis, might be corrected by chemical chaperones^{2,3}. Cystic fibrosis caused by the $\Delta F508$ mutation in the Cl^- channel protein results from chaperone-mediated retention of this mutant in the ER and subsequent degradation of this channel by the proteasome. Investigators at Vertex and a group at the University of California, San Francisco have recently identified small molecules that bind to and stabilize the $\Delta F508$ Cl^- channel in the ER, enabling the channel to exit from the ER and to be trafficked to the plasma membrane, where it exhibits sufficient activity to partly ameliorate the disease¹⁶.

$\alpha 1$ -AT is a member of the serpin family of protease inhibitors and is secreted by liver cells to control the proteolytic activity of a variety of circulating enzymes. In $\alpha 1$ -AT deficiency, a misfolded but otherwise functional mutant protein is retained in the ER of liver cells. The failure to secrete $\alpha 1$ -AT into the circulation allows neutrophil elastase to degrade lung parenchyma, causing emphysema. Perlmutter and colleagues¹⁷ have shown that osmolytes such as 4-phenylbutyric acid and glycerol increase the secretion efficiency of $\alpha 1$ -AT in cell lines and in transgenic animals. Although these compounds do not bind to $\alpha 1$ -AT specifically, they increase the secretion efficiency of variants of $\alpha 1$ -AT that are normally retained in the ER. 4-Phenylbutyric acid and glycerol accomplish this without influencing the secretion efficiency of other proteins or by decreasing proteasome activity. We speculate that these osmolytes upregulate the chaperone system in cells and in transgenic mice expressing variants of $\alpha 1$ -AT. This would explain why misfolded variants of $\alpha 1$ -AT, but not other WT proteins, are secreted more efficiently. We suspect that this is not a pure osmolytic effect in cells because the most potent osmolyte, trimethylamine N-oxide, does not increase secretion efficiency. This is reminiscent of the impact of guanidinium chloride, a potent denaturant, on curing the prion Ψ^{+} phenotype in yeast, a result attributable to chaperone changes.

Stabilizing proteins that misfold after secretion

The misfolding and misassembly that occur within the ER are largely controlled by the interplay between the chaperone and proteasome systems. Misfolding and misassembly occurring later in the secretory pathway or outside the cell, after secretion or membrane attachment, are more difficult to correct. Extracellular amyloid and related aggregates can be degraded by immune-system pathways including catabolism by macrophages. Unfortunately, aggregate clearance is often slowed down by disease and ageing, leading to the accumulation of misfolded and misassembled proteins thought to be responsible for diseases such as Alzheimer's and the familial amyloidoses.

Transthyretin amyloidosis

Stabilization of the native state is one approach to prevent post-secretory misfolding that has been studied extensively in the case of transthyretin (TTR) amyloidogenesis. TTR is secreted into the plasma as a homotetramer with 127 amino-acid subunits, and is the backup transporter for thyroxine and the primary transporter of holo-retinol-binding protein. Tetramer dissociation to a misfolded monomer is typically rate-limiting for TTR misassembly, and leads to the formation of non-fibrillar and amyloid aggregates. All of the TTR mutations associated with familial disease studied so far decrease tetramer stability and, in the most severe cases, increase the speed of dissociation and thus of amyloidogenesis¹⁸. Interallelic trans-suppression of TTR amyloidogenesis has been observed in a compound heterozygous Portuguese family, who express the disease-associated mutation V30M but do not develop familial amyloid

polyneuropathy¹⁹. These individuals make TTR tetramers composed of T119M and/or V30M subunits. The incorporation of T119M subunits imparts kinetic stability by increasing the tetramer-dissociation barrier, slowing the rate of TTR-amyloid fibril formation¹⁹. This result suggests that interallelic and intra-allelic trans-suppression of misfolding could become very powerful therapeutic strategies for the prevention of misfolding diseases. It might be difficult to use suppressor proteins delivered by exogenous supplementation owing to the slow kinetics of subunit exchange and the possible antigenicity of suppressor variants²⁰. Similar approaches have been suggested to prevent p53 misfolding in a variety of cancers²¹. Gene therapy, once practical, should be very useful for the implementation of trans-suppression.

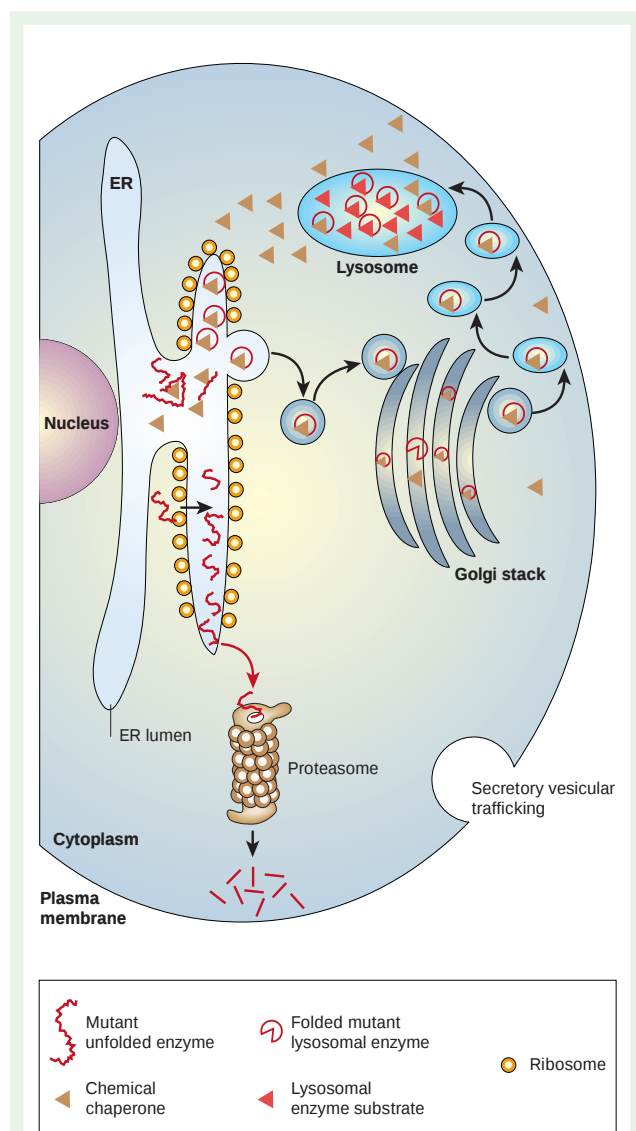


Figure 1 Folding of mutant proteins destined to be lysosomal enzymes. Cell-permeable chemical chaperones (dark red triangles) act as a template for variant lysosomal enzymes as they fold in the ER. Some of these lysosomal enzymes would otherwise misfold and be retained in the ER and ultimately be degraded by the proteasome. Chemical chaperones facilitate folding in the ER lumen, which is necessary for proper trafficking to the lysosome where a high concentration of substrate (red triangles) exists to compete with the chemical chaperones that bind to the active site. Although the chemical chaperones are crucial for folding at pH 7 (for example, in the ER), the variant lysosomal enzymes might not need the chaperones when they arrive in the lysosome (where the pH is about 5) because their sequences are stably folded at this pH.

Evidence that trans-suppression ameliorates TTR amyloid disease provides both an incentive and the biological rationale to search for small molecules that stabilize the native state and/or destabilize the transition state associated with TTR amyloidogenesis²². Screening and structure-based design have identified six distinct classes of small molecules that bind with high affinity to the unoccupied thyroxine-binding sites within TTR. These small molecules stabilize the native state, preventing amyloidogenesis and denaturation at low pH and in a wide variety of chaotropes and organic solvents for extended periods (for example, weeks)¹⁹. When ligand binding stabilizes the native state by more than 3 kcal mol⁻¹ (about 12.5 kJ mol⁻¹), the concentration of the misfolded amyloidogenic monomer decreases by more than 100-fold (assuming that the barriers associated with misfolding are readily surmountable). Decreasing the concentration of the aggregation-prone species is important because amyloidogenesis is a high-order kinetic process; decreasing the concentration modestly can therefore markedly influence the aggregation rate²³. Stabilization of the native state by a ligand also has kinetic ramifications when the ligand selectively stabilizes the native state over the transition state¹⁹. TTR tetramer stabilization by ligand binding often increases the $\Delta G^\ddagger$ of dissociation (see Box 1), markedly slowing the typical rate-limiting step of TTR amyloidogenesis. The mechanistic similarities between trans-suppression and native-state small-molecule stabilization imply that altering denaturation energetics with ligands will probably be effective in humans.

Prion disease

The agreement between the various mouse models of prion disease and their human and bovine counterparts provides a prototypic setting for the exploration of therapeutic approaches to diseases of misfolding. These therapeutic efforts can be divided into two broad classes, serendipitous and mechanistic. A variety of compounds including anthracyclines, porphyrins and diazo dyes block prion replication when administered with PrP^{Sc}, the aggregated infectious form of the prion protein, into animal models^{24,25}. Unfortunately, this is not a clinically relevant model for therapeutic intervention, because subclinical disease exists for months in mice and years in humans. Efforts to identify more potent analogues of these molecules have been disappointing; however, quinacrine might prove to be a different case. As quinacrine is an approved pharmaceutical, clinical trials are being carried out to test the usefulness of this molecule in patients with Creutzfeldt-Jakob disease. Medicinal chemistry efforts are underway to identify more potent agents. For example, some bisacridines are 10–20-fold more potent than quinacrine. Animal studies are underway to characterize the efficacy of bisacridines in the treatment of prion disease. Recently, Zahn and colleagues²⁶ have used NMR spectroscopy and chemical-shift titration to show that quinacrine binds specifically to PrP^C, the normally folded cellular isoform of the prion protein. This observation could accelerate drug-development efforts.

Studies of heterozygous and homozygous mice with PrP knocked out show that decreasing or eliminating PrP^C production will slow or prevent prion infection. Similarly, antibody studies suggest that more efficient clearance of PrP could slow or cure prion disease^{27–29}. These results complement similar efforts with Alzheimer's disease³⁰. As with other aggregation-prone proteins, the tendency for PrP^C to misfold can be altered by proteins that bind to PrP^C. In prion disease, PrP^{Sc} clearly induces misfolding, so antibodies or small molecules that antagonize this interaction should prove useful. There is also genetic evidence that other host molecules facilitate PrP^{Sc} replication³¹. Dominant-negative variants of PrP^C slow or stop PrP^{Sc} replication in cell culture and in transgenic animals. Small molecules designed to mimic this dominant-negative effect can block PrP^{Sc} replication in cell culture as well.

Multiple myeloma

Immunoglobulin light-chain amyloidoses are particularly challenging from a therapeutic perspective. Multiple myeloma (plasma-cell

cancer) or plasma-cell dyscrasia overproduce an antibody light-chain clone that is prone to aggregation. Light-chain aggregation often results in progressive organ failure, making it difficult to use autologous blood stem-cell transplantation to treat the cancerous component of this disease. Direct approaches to decrease light-chain production by the use of specific antibodies, antisense RNA or short interfering RNA are likely to be difficult owing to the patient-specific features of each plasma-cell dyscrasia³². However, it is possible that Velcade, a proteasome inhibitor that is effective in the treatment of multiple myeloma, could have a paradoxical palliative effect on light-chain aggregation. Plasma cells need an efficient proteasome system to manipulate antibodies, so proteasome inhibitors could disable or kill neoplastic cells faster than new cells arise, decreasing the manufacture of light chains.

Intervention in Alzheimer's disease

The risk of Alzheimer's disease is strongly influenced by two competitive ER processing pathways. The amyloid precursor protein (APP) has a single membrane-spanning helix that is processed by either α -secretase (about 90%) or by the β - and γ -secretases (about 10%). The cleavage by α -secretase produces two fragments that do not produce amyloids. The β -secretase and subsequent γ -secretase cleavages yield three fragments; the smallest of which is a family of 39–43-residue peptides referred to as A β . These fragments have a tendency to assemble into a variety of aggregates, including amyloid fibrils. Generally speaking, the longer the A β peptide, the more amyloidogenic it is. Its exact length distribution is controlled by the allosteric modulation of γ -secretase.

Inhibition of A β -fibril formation might be a reasonable therapeutic strategy, because familial mutations that lead to an increase in A β concentration or to its aggregation increase neuropathology^{33–35}. Peptidomimetics, based on the peptide LVFFA from A β , modified at the amino or carboxy terminus, and the all-D (right-handed) version and several retroinverso peptidomimetics, block both A β seeding and growth. Unfortunately the pharmacological profile of these compounds is far from ideal³³. Non-peptidic, aromatic-rich A β aggregation inhibitors have also been reported by many pharmaceutical and biotechnology companies. These compounds typically inhibit the aggregation of A β and several other amyloidogenic polypeptides. Their lack of specificity might be a problem; it now seems likely that mammalian melanocytes and related cell types make amyloid for functional purposes³⁶. Moreover, recent evidence suggests that soluble, diffusible non-fibrillar A β aggregates lead to pathology in Alzheimer's disease^{5,37–39}. This calls into question the merits of the subset of these compounds that shift the distribution from fibrils towards soluble, diffusible non-fibrillar A β aggregates.

Because A β aggregation is highly dependent on monomer concentration (because of its high kinetic order) and on the distribution of A β family members, substantial effort has gone into the development of inhibitors, activators and modulators of the secretases^{40,41}. Decreasing the concentration of A β below the critical concentration prevents aggregation⁴², and so it is not necessary to block all A β production. Several β -secretase inhibitors are now available, and this seems to be a promising approach because mice lacking β -secretase do not have an overt phenotype. Considerable progress has also been made in developing γ -secretase inhibitors. However, the side-effect profile for these inhibitors must be considered carefully: mice lacking γ -secretase die *in utero* because this enzyme processes notch, a regulatory protein required for development. A therapeutic window for γ -secretase inhibitors will require substantial inhibition of A β processing without the complete inhibition of notch processing. Golde and Koo⁴³ have shown that some non-steroidal anti-inflammatory drugs can influence the processing of APP to produce less A β 1–42; others have shown that cholesterol depletion can markedly decrease the overall amount of A β produced.

When APP transgenic mice overexpressing A β are actively immunized with A β peptides, deposits that would ordinarily accumulate

are cleared⁴⁴. Passive immunotherapy with intravenous A β antibody also prevents the formation of A β deposits in animals that would otherwise be deposit-laden. Unfortunately, active immunization of patients with Alzheimer's disease using A β 1–42 revealed antibodies crossing the blood–brain barrier causing meningoencephalitis-like inflammation in 4% of the trial participants, presumably originating from microglial activation⁴⁵. This resulted in the discontinuation of the clinical trial. Cautious optimism still exists for passive immunotherapy in humans, because anti-A β antibody infusion could be stopped if meningoencephalitis symptoms develop⁴⁶.

Future considerations

Although thermodynamics undoubtedly has a role in misfolding diseases, it is the kinetics of misfolding and misassembly that seems to have the dominant influence. New therapeutics are likely to come from medicinal chemistry efforts and clinical trial designs based on the considerations discussed here. We therefore expect that orally administered chemical chaperones that treat lysosomal storage diseases will replace intravenous enzyme replacement therapy as the standard treatment for these rare inherited conditions. Some drugs have already been identified that block the destruction of mutant integral membrane proteins exhibiting suboptimal folding in the ER. We believe that therapeutics based on similar approaches will soon enter clinical trials for patients with cystic fibrosis and α 1-AT deficiency. For the amyloid-deposition diseases, small alterations in the rate of aggregation mediated by stabilization of the native state or depletion of the precursor could stabilize or improve patients with these diseases. Enhanced clearance of amyloidogenic aggregates by immunotherapy also seems promising. Given the challenges posed by unregulated immune responses after vaccination, passive immunotherapy is more likely to be available in the short term. More clinically palatable small-molecule approaches will follow with the more typical timeline for pharmaceutical discovery and development. The histological aggregates observed by Lewy, Pick and other noted pathologists more than a century ago have recently been identified as largely homogeneous deposits of misfolded proteins. A large and increasing number of diseases are now associated with the accelerated production and/or delayed clearance of protein aggregates. With the development of pharmacological approaches to these diseases, we are participating in a journey from the bedside to the bench and now back to the bedside. □

doi:10.1038/nature02265

- Anfinsen, C. B. Principles that govern the folding of protein chains. *Science* **181**, 223–230 (1973).
- Powell, K. & Zeitlin, P. L. Therapeutic approaches to repair defects in Δ F508 CFTR folding and cellular targeting. *Adv. Drug Deliv. Rev.* **54**, 1395–1408 (2002).
- Howard, M. & Welch, W. J. Manipulating the folding pathway of Δ F508 CFTR using chemical chaperones. *Methods Mol. Med.* **70**, 267–275 (2002).
- Carrell, R. W. & Lomas, D. A. Mechanisms of disease. Alpha1-antitrypsin deficiency—a model for conformational diseases. *N. Engl. J. Med.* **346**, 45–53 (2002).
- Rochet, J.-C. & Lansbury, P. T. Jr. Amyloid fibrillogenesis: themes and variations. *Curr. Opin. Struct. Biol.* **10**, 60–68 (2000).
- Kelly, J. W. Alternative conformations of amyloidogenic proteins govern their behavior. *Curr. Opin. Struct. Biol.* **6**, 11–17 (1996).
- Fan, J.-Q. A contradictory treatment for lysosomal storage disorders: inhibitors enhance mutant enzyme activity. *Trends Pharm. Sci.* **24**, 355–360 (2003).
- Sawkar, A. R. *et al.* Chemical chaperones increase the cellular activity of N370S β -glucosidase: a therapeutic strategy for Gaucher disease. *Proc. Natl Acad. Sci. USA* **99**, 15428–15433 (2002).
- Morello, J.-P. *et al.* Pharmacological chaperones rescue cell-surface expression and function of misfolded V2 vasopressin receptor mutants. *J. Clin. Inv.* **105**, 887–895 (2000).
- Petaja-Repo, U. E. *et al.* Ligands act as pharmacological chaperones and increase the efficiency of δ opioid receptor maturation. *EMBO J.* **21**, 1628–1637 (2002).
- Janovick, J. A. *et al.* Structure–activity relations of successful pharmacologic chaperones for rescue of naturally occurring and manufactured mutants of the gonadotropin-releasing hormone receptor. *J. Pharm. Exp. Ther.* **305**, 608–614 (2003).
- Noorwez, S. M. *et al.* Pharmacological chaperone-mediated *in vivo* folding and stabilization of the P23H-opsin mutant associated with autosomal dominant retinitis pigmentosa. *J. Biol. Chem.* **278**, 14442–14450 (2003).
- Ficker, E., Obejero-Paz, C. A., Zhao, S. & Brown, A. M. The binding site for channel blockers that rescue misprocessed human long QT syndrome type 2 ether-a-gogo-related gene (HERG) mutations. *J. Biol. Chem.* **277**, 4989–4998 (2002).
- Fan, J.-Q., Ishii, S., Asano, N. & Suzuki, Y. Accelerated transport and maturation of lysosomal α -galactosidase A in Fabry lymphoblasts by an enzyme inhibitor. *Nature Med.* **5**, 112–115 (1999).
- Frustaci, A. *et al.* Improvement in cardiac function in the cardiac variant of Fabry's disease with galactose-infusion therapy. *N. Engl. J. Med.* **345**, 25–32 (2001).
- Springsteel, M. F. *et al.* Benzoflavone activators of the cystic fibrosis transmembrane conductance regulator: towards a pharmacophore model for the nucleotide-binding domain. *Bioorg. Medic. Chem.* **11**, 4113–4120 (2003).
- Burrows, J. A. J., Willis, L. K. & Perlmutter, D. H. Chemical chaperones mediate increased secretion of mutant α 1-antitrypsin (α 1-AT) Z: a potential pharmacological strategy for prevention of liver injury and emphysema in α 1-AT deficiency. *Proc. Natl Acad. Sci. USA* **97**, 1796–1801 (2000).
- Hammarstrom, P., Jiang, X., Hurshman, A. R., Powers, E. T. & Kelly, J. W. Sequence-dependent denaturation energetics: a major determinant in amyloid disease diversity. *Proc. Natl Acad. Sci. USA* **99**, 16427–16432 (2002).
- Hammarstrom, P., Wiseman, R. L., Powers, E. T. & Kelly, J. W. Prevention of transthyretin amyloid disease by changing protein misfolding energetics. *Science* **299**, 713–716 (2003).
- Schneider, F., Hammarstrom, P. & Kelly, J. W. Transthyretin slowly exchanges subunits under physiological conditions: a convenient chromatographic method to study subunit exchange in oligomeric proteins. *Protein Sci.* **10**, 1606–1613 (2001).
- Nikolova, P. V., Wong, K.-B., DeDecker, B., Henckel, J. & Fersht, A. R. Mechanism of rescue of common p53 cancer mutations by second-site suppressor mutations. *EMBO J.* **19**, 370–378 (2000).
- Sacchettini, J. C. & Kelly, J. W. Therapeutic strategies for human amyloid diseases. *Nature Rev. Drug Discov.* **1**, 267–275 (2002).
- White, J. T. & Kelly, J. W. Support for the multigenic hypothesis of amyloidosis: the binding stoichiometry of retinol-binding protein, vitamin A, and thyroid hormone influences transthyretin amyloidogenicity *in vitro*. *Proc. Natl Acad. Sci. USA* **98**, 13019–13024 (2001).
- Korth, C., May, B. C. H., Cohen, F. E. & Prusiner, S. B. Acridine and phenothiazine derivatives as pharmacotherapeutics for prion disease. *Proc. Natl Acad. Sci. USA* **98**, 9836–9841 (2001).
- May, B. C. H. *et al.* Potent inhibition of scrapie prion replication in cultured cells by bis-acridines. *Proc. Natl Acad. Sci. USA* **100**, 3416–3421 (2003).
- Voghterr, M. *et al.* Antimalarial drug quinacrine binds to C-terminal helix of cellular prion protein. *J. Med. Chem.* **46**, 3563–3564 (2003).
- Peretz, D. *et al.* Antibodies inhibit prion propagation and clear cell cultures of prion infectivity. *Nature* **412**, 739–743 (2001).
- White, A. R. *et al.* Monoclonal antibodies inhibit prion replication and delay the development of prion disease. *Nature* **422**, 80–83 (2003).
- Heppner, F. L. *et al.* Prevention of scrapie pathogenesis by transgenic expression of anti-prion protein antibodies. *Science* **294**, 178–182 (2001).
- Solomon, B. Anti-aggregating antibodies, a new approach towards treatment of conformational diseases. *Curr. Med. Chem.* **9**, 1737–1749 (2002).
- Telling, G. C. *et al.* Prion propagation in mice expressing human and chimeric PrP transgenes implicates the interaction of cellular PrP with another protein. *Cell* **83**, 79–90 (1995).
- Ohno, S. *et al.* The antisense approach in amyloid light chain amyloidosis: identification of monoclonal Ig and inhibition of its production by antisense oligonucleotides *in vitro* and *in vivo* models. *J. Immunol.* **169**, 4039–4045 (2002).
- Findeis, M. A. Peptide inhibitors of beta amyloid aggregation. *Curr. Top. Med. Chem.* **2**, 417–423 (2002).
- Yang, D.-S. *et al.* Assembly of Alzheimer's amyloid- β fibrils and approaches for therapeutic intervention. *Amyloid* **8**, 10–19 (2001).
- Howlett, D. R. *et al.* Inhibition of fibril formation in β -amyloid peptide by a novel series of benzofurans. *Biochem. J.* **340**, 283–289 (1999).
- Kelly, J. W. & Balch, W. E. Amyloid as a natural product. *J. Cell Biol.* **161**, 461–462 (2003).
- Lamber, M. P. *et al.* Diffusible, nonfibrillar ligands derived from A β 1–42 are potent central nervous system neurotoxins. *Proc. Natl Acad. Sci. USA* **95**, 6448–6453 (1998).
- Kayed, R. *et al.* Common structure of soluble amyloid oligomers implies common mechanism of pathogenesis. *Science* **300**, 486–489 (2003).
- Walsh, D. M. *et al.* Naturally secreted oligomers of amyloid β protein potently inhibit hippocampal long-term potentiation *in vivo*. *Nature* **416**, 535–539 (2002).
- Wolfe, M. S. γ -Secretase as a target for Alzheimer's disease. *Curr. Top. Med. Chem.* **2**, 371–383 (2002).
- Vassar, R. β -Secretase (BACE) as a drug target for Alzheimer's disease. *Adv. Drug Deliv. Rev.* **54**, 1589–1602 (2002).
- Jarrett, J. T. & Lansbury, P. T. Jr. Seeding 'one-dimensional crystallization' of amyloid: A pathogenic mechanism in Alzheimer's disease and scrapie? *Cell* **73**, 1055–1058 (1993).
- Golde, T. E., Eriksen, J. L., Weggen, S., Sagi, S. A. & Koo, E. H. Nonsteroidal antiinflammatory drugs as therapeutic agents for Alzheimer's disease. *Drug Dev. Res.* **56**, 415–420 (2002).
- Spooner, E. T., Desai, R. V., Mori, C., Leverone, J. F. & Lemere, C. A. The generation and characterization of potentially therapeutic A β antibodies in mice: differences according to strain and immunization protocol. *Vaccine* **21**, 290–297 (2002).
- Nicoll, J. A. R. *et al.* Neuropathology of human Alzheimer disease after immunization with amyloid- β peptide: a case report. *Nature Med.* **9**, 448–452 (2003).
- Solomon, B. Protective molecules in Alzheimer's disease: therapeutic antibodies. *Drug News Perspect.* **15**, 410–416 (2002).
- Colom, W. & Kelly, J. W. Partial denaturation of transthyretin is sufficient for amyloid fibril formation *in vitro*. *Biochemistry* **31**, 8654–8660 (1992).
- Fandrich, M., Fletcher, M. A. & Dobson, C. M. Amyloid fibrils from muscle myoglobin. *Nature* **410**, 165–166 (2001).
- Baskakov, I. V., Legname, G., Prusiner, S. B. & Cohen, F. E. Folding of prion protein to its native α -helical conformation is under kinetic control. *J. Biol. Chem.* **276**, 19687–19690 (2001).

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs Sales Director:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Netherlands/ Italy/ Spain/

Portugal/ Belgium:

Evelina Rubio Hakansson (4973)

Scandinavia: Sille Opstrup (4994)

France/ Switzerland:

Amelie Pequignot (4974)

Natureevents:

Paul Constant (4954)

Production Manager:

Billie Franklin
To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Office

Germany/ Austria:

Patrick Phelan, Odo Wulffen

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707

Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: naturejobs@natureny.com

US Sales Manager:

Peter Bless

US Advertising Coordinator:

Linda Adam

Japan Head Office, Tokyo

MG Ichigaya Building (5F),

19-1 Harajukutamachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Rinoko Asami

e-mail: rasami@naturejpn.com

naturejobs

Conference management

With heightened security, travel over the past couple of years has become more taxing. And with tightened budgets, it has also become less affordable. So scientific travel — whether for workshops, conferences or symposia — requires greater justification. The best way to spend a shrinking travel budget is to get more from fewer events. And the best way to do that is to set yourself some goals and then do some planning.

As a journalist, when I attend scientific conferences, I try to come away with at least one short item I can use immediately (for instance, some fodder for this column), enough information for a longer feature that could run later in the year, and an idea or contact that will develop into a long-term project. I intend to travel less next year, although when I do attend events or make lab visits, I also intend to bring back a greater number of contact names. In addition, I'm planning to have a more concrete idea of future interactions with these contacts before I leave the event, and to follow up faster on the meeting after I return.

If I were a working scientist, I would make similar goals, such as sitting in on at least one session about which I know absolutely nothing, forcing myself to meet five new people outside my normal community, or trying to pick up at least one new skill. If I were a group leader, I would try to find some more potential postdocs, students and collaborators — and, perhaps, one potential multidisciplinary partner from outside my area of expertise.

What is the best way to make all of this happen? Simple: start planning now. Many courses fill up fast and are oversubscribed. And picking a conference to attend on the basis of the fairly broad names used to describe many sessions requires considerable skill. We hope that the events directory in this issue and the natureevents.com website will help you through the process.

Paul Smaglik
Naturejobs editor



Contents

EVENTS DIRECTORY 2004

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

ANNOUNCEMENTS

EVENTS